マスクされた単語の埋め込みと 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

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@c.mbox.nagoya-u.ac.jp　{sasano,takedasu}@i.nagoya-u.ac.jp

1 はじめに

本研究では，テキスト中の動詞が喚起する意味フレームの推定に取り組む．具体的には，FrameNet [1] で定義されているフレームごとにテキスト中の動詞をクラスタリングすることを目標とする．たとえば，表 1 の (1)〜(4) に示す FrameNet の用例の場合，動詞が喚起するフレームごとに{(1)}, {(2)}, {(3), (4)}の 3 つのクラスタにまとめることが目標となる．

動詞の意味フレーム推定では，近年 ELMo [2] や BERT [3] などの文脈化単語埋め込みの有用性が報告されている．たとえば，SemEval2019 の共通タスク [4] でベースラインを超えた 3 手法 [5, 6, 7] はいずれもフレーム推定対象動詞の文脈化単語埋め込みを利用したクラスタリングに基づく手法となっている．しかし，これらの手法には 2 つの問題点がある．

1 つ目の問題点は，推定対象となる動詞の表層的な情報の影響が大きいことである．表 1 の「get」のように，一部の動詞は文脈により異なるフレームを喚起する．しかし，文脈化単語埋め込みは対象単語の表層的な情報も含むことから，同一の動詞の埋め込みは類似する傾向がある．このため，クラスタリングの結果，同一の動詞の用例が 1 つのクラスタにまとめられることが多い．たとえば，FrameNet から「get」と「acquire」の用例を抽出し，事前学習済みの BERT を用いてこれらの文脈化単語埋め込みを獲得し，t-SNE[8] で 2 次元にマッピングした結果を図 1 左に示す．「get」が喚起するフレームのうち，Getting フレームが付与された用例の動詞の文脈化単語埋め込みは，同じく Getting フレームを喚起する「acquire」の用例の動詞の文脈化単語埋め込みに近い位置に分布する傾向はあるものの，喚起するフレーム間の差より動詞間の差の方が大きいことが確認できる．

本研究ではこの問題を解消するため，対象動詞をマスクした文脈化単語埋め込みを利用する手法を提案する．図 1 左と同様に，マスクされた動詞の文脈化単語埋め込みを 2 次元にマッピングした結果を図 1 右に示す．マスクを用いた場合は，動詞の表層的な情報は限定的となり，同一のフレームを喚起する用例が近い位置に分布することが確認できる．

2 つ目の問題点は，全動詞の用例を一度にまとめてクラスタリングしていることである．この結果，各動詞が喚起するフレームの異なり数は多くても数個程度と限定的であるにも関わらず，1 つの動詞の用例が多くの異なるクラスタに属すると判断される可能性がある．本研究では，このような事態を避けるため，まず動詞ごとに用例のクラスタリングを行った後，動詞横断的にフレーム単位でまとめる 2 段階クラスタリングに基づく手法を提案する．

表 1: FrameNet の動詞「get」と「acquire」の用例と各動詞が喚起するフレーム (括弧内).

(1)	We'll not get there before the rain comes.	(Arriving)
(2)	The problem continued to get worse.	(Transition_to_state)
(3)	You may get more money from the basic pension.	(Getting)
(4)	We have acquired more than 100 works.	(Getting)

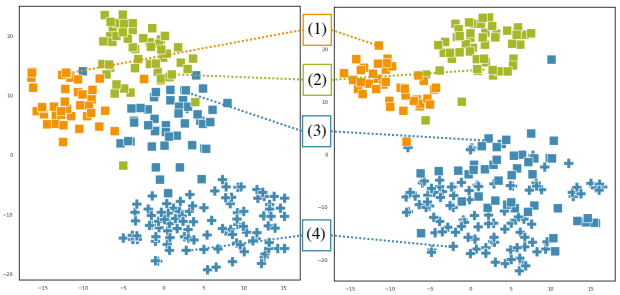

図 1: 動詞「get」と「acquire」の動詞 (左) とマスクされた動詞 (右) の BERT による埋め込みの 2 次元マッピング．番号は表 1 に対応し，■ と + は動詞「get」と「acquire」，各色は Arriving，Transition_to_state，Getting フレームを示す．

2 提案手法

本研究では，テキスト中の動詞が喚起するフレームの推定に，マスクされた単語の埋め込み，および，2 段階クラスタリングを利用する手法を提案する．

2.1 マスクされた単語の埋め込みの利用

提案手法では動詞のフレーム推定に利用する埋め込みとして，従来手法で使用された推定対象動詞の文脈化単語埋め込みに加え，その動詞をマスクした場合の文脈化単語埋め込みを利用する．本研究では以下の 3 種類の文脈化単語埋め込みを考える．

1. v_{WORD}: 対象の動詞の通常の文脈化単語埋め込み
2. v_{MASK}: 対象の動詞を “[MASK]” に置き換えた場合の文脈化単語埋め込み
3. $v_{\text{W+M}}$: 次式で定義される上記 2 つの加重平均

$$v_{\text{W+M}} = (1-\alpha)\cdot v_{\text{WORD}} + \alpha \cdot v_{\text{MASK}} \qquad (1)$$

$v_{\text{W+M}}$ は対象の動詞をマスクした場合と，しない場合の文脈化単語埋め込みの加重平均である．開発セットを用いて重み α を適切に設定することにより，対象の動詞の表層的な情報の重みと周辺文脈から得られる情報の重みを適切に考慮した埋め込みが得られることを期待している．α を 0 とした場合は v_{WORD} と，1 とした場合は v_{MASK} と一致する．

2.2 2 段階クラスタリング

提案手法では，1 段階目で動詞ごとに用例のクラスタリングを行った後，2 段階目で動詞横断的なクラスタリングを行い，最終的に生成されたクラスタがそれぞれ 1 つのフレームに対応すると考える．1 段階目のクラスタリングにおいて，各動詞の用例を少数のクラスタにまとめることにより，1 つの動詞の用例が多くの異なるクラスタに属することが避けられることを期待している．

図 2 に「get」と「acquire」の用例を 2 段階クラスタリングしたときの流れを示す．この例では，1 段階目のクラスタリングの結果，「get」の用例は 3 つ，「acquire」の用例は 1 つのクラスタにまとめられ，2 段階目のクラスタリングにより，「get」のクラスタの 1 つと「acquire」のクラスタがマージされ，最終的に 3 つのクラスタにまとめられている．以下では，各クラスタリング手法の詳細について説明する．

動詞ごとの用例クラスタリング　1 段階目のクラスタリングは，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ることを目的とする．クラスタリング対象の用例の動詞は共通であり，文脈化単語埋め込みとして v_{WORD} を用いる場合と，v_{MASK} を用いる場合で結果に大きな違いはないと考えられることから，1 段階目のクラスタリングでは文脈化単語

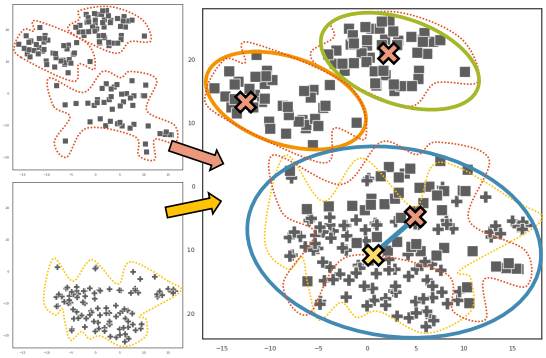

図 2: 2 段階クラスタリングの流れ．左上，左下図はそれぞれ「get」と「acquire」を対象とした 1 段階目のクラスタリング，右図は 2 段階目のクラスタリングを示している．各図における ■ と + はそれぞれ「get」と「acuquire」の各用例の埋め込みを表す．

埋め込みとして v_{MASK} のみを用いる．クラスタリング手法としては，X-means [9]，または，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用いる．ここで，X-means は自動でクラスタ数を決定する手法であるが，階層型クラスタリングはクラスタリングの終了基準を必要とする．群平均法では，クラスタ間距離をクラスタをまたがる全要素ペアの平均距離により定義し，クラスタ間距離が小さいクラスタから順にマージするが，本研究ではクラスタ間距離が閾値 θ 以下となるクラスタペアがなくなった時点でクラスタリングを終了する．閾値 θ は全動詞で共有され，十分に大きな値に設定した場合，すべての動詞についてクラスタは 1 つとなる．本研究では，θ を十分に大きな値から徐々に小さくしていき，全動詞のクラスタ数の平均が，開発セットにおいて各動詞が喚起するフレーム数の平均と一致する値に設定する．

フレーム意味論では，語と意味フレームを結び付けたものを語彙項目 (Lexical Unit; LU) と呼ぶ．1 段階目のクラスタリングにより生成されたクラスタは，各動詞の用例をそれが喚起するフレームごとにまとめたものであることから，各クラスタは LU に対応する集合とみなせることができ，本稿では 1 段階目のクラスタリングで生成された各クラスタを疑似 LU (pseudo-LU; pLU) と呼ぶことにする．

動詞横断的クラスタリング　2 段階目のクラスタリングでは，1 段階目のクラスタリングで生成した pLU を，動詞横断的にそれが喚起するフレームごとにまとめることを目的とする．まず，pLU ごとに各用例の文脈化単語埋め込みの平均を算出し，その後，算出された平均埋め込みを用いて動詞横断的に

表 2: FrameNet から作成したデータセット

	#動詞	#LU	#フレーム	#用例
開発セット	255	300	169	12,718
テストセット	1,017	1,188	393	47,499
全体	1,272	1,488	434	60,217

クラスタリングを行う．具体的なクラスタリング手法には，ユークリッド距離に基づく群平均法あるいはウォード法による階層型クラスタリングを用いる．

クラスタリングは，2 つの pLU が同じクラスタに属する割合 $P_{C_1=C_2}$ が，開発セットにおいて 2 つの LU が同じフレームに属する割合 $P_{F_1=F_2}$ 以上となった時点で終了する．ここで，$P_{F_1=F_2}$ は式 (2) により算出される．一方，$P_{C_1=C_2}$ も同様に計算できるが，pLU の全ペア数はクラスタリングの段階に依らず一定であるのに対し，同じクラスタに属する pLU のペア数はクラスタリングが進むにつれて単調増加し，全体が 1 つのクラスタとなった時点で 1 となる．このため，クラスタリングの過程で $P_{F_1=F_2}$ 以上の値となることが保証される．また，無作為に抽出した 2 つの LU が同じフレームに属する確率はデータサイズに依らないことから，このような基準はテストセットのサイズに依らず有効であると考えられる．

$$p_{F_1=F_2} = \frac{\text{同じフレームに属する LU のペア数}}{\text{LU の全ペア数}} \quad (2)$$

3 実験

提案手法の有効性を確認するため，テキスト中の動詞の意味フレーム推定実験を行った．

3.1 実験設定

データセット FrameNet[1]から，いずれかのフレームにおいて 20 以上の用例をもつ LU となっている動詞，および，該当する LU の用例を抽出し実験に使用した．LU ごとの用例数は最大 100 件とし，100 件を越える場合は無作為に 100 件を選択し用いた．抽出された動詞は全部で 1,272 個であり，そのうち 20%の 255 動詞を開発セットとして，残りの 1,017 動詞をテストセットとして使用した．この際，複数のフレームを喚起する動詞[2]の割合が開発セットとテストセットで一致するように留意した．なお，開発セットは各種パラメータ，および，文脈化単語埋め込みとして使用する層の決定に利用した．表 2 に作成したデータセットの統計値を示す．

1 https://framenet.icsi.berkeley.edu/fndrupal/
2 約 14%にあたる 178 動詞が該当した．

比較モデル 提案手法では，1 段階目のクラスタリング法として群平均法による階層型クラスタリングまたは X-means を，2 段階目のクラスタリング法としてウォード法または群平均法による階層型クラスタリングを用いることから，提案モデルとしてこれらを組み合わせた計 4 種類のモデルを比較した．また，1 動詞をそのまま 1 クラスタとして扱うモデル (1-cluster-per-verb; 1cpv)，1 動詞を 1 クラスタ (1cpv') にまとめた上で 2 段階目のクラスタリングを行うモデルとも比較した．

SemEval2019 の共通タスクのスコアの高かった上位 3 つの先行モデルとの比較も行った．Arefyev ら [5] は，推定対象動詞の BERT の埋め込みを用いてコサイン類似度に基づく群平均法による階層型クラスタリングをした後，BERT を用いて得た推定対象動詞の言い換え単語による TF-IDF により特徴量を生成して，各クラスタを 2 分するクラスタリングを行っている．Anwar ら [6] は，skip-gram [10] で推定対象動詞の埋め込みと文全体の埋め込みを連結した表現を用いてマンハッタン距離に基づく群平均法による階層型クラスタリングを行っている．Ribeiro ら [7] は，推定対象動詞の ELMo の埋め込みを獲得して Chinese Whispers [11] によるグラフクラスタリングを行っている．

2 段階クラスタリングの有用性を確認するため，1 段階でクラスタリングを行うモデルとの比較も行った．1 段階クラスタリングに基づくモデルでは，提案モデルと同様に，文脈化単語埋め込みとして v_{W+M} を使用し，重み α は開発セットを用いて調整し，ウォード法または群平均法による階層型クラスタリングを用いた．クラスタ数に関しては正解のフレーム数を人手で与え，クラスタ数が正解のフレーム数と一致した時点でクラスタリングを停止した．

実験設定 評価尺度として，B-Cubed Precision (BcP)，B-Cubed Recall (BcR)，およびその調和平均である B-Cubed F-score (BcF) と，Purity (Pu)，Inverse Purity (iPu)，およびその調和平均である F-score (PiF) の 6 つの尺度を利用した．B-Cubed はクラスタ集合と人手で付与されたフレーム集合の用例の分布に着目した用例単位の評価指標，Pu と iPu はそれぞれクラスタ内のフレームの一貫性と同一フレームによるクラスタの集中性を評価する指標である．また，文脈化単語埋め込みは，Hugging Face が公開している Transformers[3]に含まれる事前学習済みの BERT

3 https://github.com/huggingface/transformers

表 3: フレーム推定実験の結果．#pLU は 1 段階目のクラスタリング後の pLU 数，#C はクラスタ数を表す．

モデル	クラスタリング法		α	#pLU	#C	Pu / ıPu / PıF	BcP / BcR / BcF
1 動詞 1 クラスタ	1cpv		-	-	1017	88.9 / 39.7 / 54.9	86.6 / 33.9 / 48.7
Arefyev et al. (2019) [5]	群平均法 (コサイン類似度)		-	-	995	69.9 / 55.1 / 61.6	62.8 / 44.0 / 51.7
Anwar et al. (2019) [6]	群平均法 (マンハッタン距離)		-	-	891	71.5 / 52.0 / 60.2	65.1 / 41.0 / 50.3
Ribeiro et al. (2019) [7]	Chinese Whispers		-	-	542	50.9 / 66.3 / 57.5	39.4 / 56.7 / 46.5
1 段階	ウォード法		0.0	-	393	64.3 / 49.5 / 56.0	55.2 / 38.9 / 45.6
クラスタリング	群平均法		0.0	-	393	38.7 / 64.9 / 48.5	26.1 / 52.5 / 34.9
	1 段階目	**2 段階目**					
	1cpv’	ウォード法	0.8	1017	164	54.8 / 73.1 / 62.7	43.1 / 64.3 / 51.6
	1cpv’	群平均法	0.9	1017	412	69.0 / 71.3 / 70.1	60.5 / 62.3 / 61.4
2 段階	群平均法	ウォード法	0.9	1196	291	49.3 / 72.9 / 58.8	37.3 / 64.6 / 47.3
クラスタリング	群平均法	群平均法	0.6	1196	479	63.0 / **76.3** / 69.0	52.8 / **68.0** / 59.4
	X-means	ウォード法	0.8	1043	167	54.0 / 72.2 / 61.8	42.6 / 63.6 / 51.1
	X-means	群平均法	0.7	1043	410	**71.9** / 74.1 / **73.0**	**63.2** / 65.5 / **64.4**

(bert-base-uncased) を利用した．

3.2 実験結果

表 3 に FrameNet データセットを用いた実験の結果を示す．SemEval2019 の共通タスクにおいてシステムの順位付けに使用された BcF で比較した場合，1 段階目に X-means，2 段階目に群平均法によるクラスタリングを行う提案モデルが 64.4 となり，全手法の中で最も高いスコアを達成した．PıF においても最も高いスコアを達成するなど，すべての指標で高いスコアとなっている．また，2 段階目のクラスタリングの終了基準に関しても，正解のフレーム数は 393 であるのに対してクラスタ数は 410 であり，有効に機能していることが確認できる．

マスクされた単語の埋め込みの有用性　全ての 2 段階クラスタリングに基づく手法において α は 0.0, 1.0 以外の値となっていることから，v_{WORD} と v_{MASK} の両方を考慮することが有効であることが確認できる．また，いずれの手法においても α は 1.0 に近い値となっていることから，動詞横断的クラスタリングにおいては v_{MASK} の方が有用であると考えられる．一方，1 段階クラスタリングに基づく手法における $v_{\text{W+M}}$ は $\alpha = 0.0$ で v_{WORD} と同じ埋め込みとなっており，用例全体でクラスタリングする際の v_{MASK} の有用性は確認できなかった．

2 段階クラスタリングの有用性　2 段階目のクラスタリング手法として群平均法を用いた場合に全体的に高いスコアとなっており，2 段階クラスタリングの有用性が確認できる．また，1 動詞をそのまま 1 クラスタとして扱う (1cpv) よりも，1 クラスタにまとめた (1cpv’) 後に動詞横断的にクラスタリングを行う 2 段階手法の方が高いスコアとなった．さらに，1 段階目で pLU を生成する際も，1cpv’ の後に群平均法によるクラスタリングを利用するモデルよりも X-means の後に群平均法によるクラスタリングを利用したモデルが高いスコアとなっており，動詞ごとの用例クラスタリングも有用であることが確認できる．

ここで，2 段階目のウォード法によるクラスタリングが群平均法によるクラスタリングより全体的にスコアが低いのは，ウォード法がクラスタ内の用例数が増えるほど結合しづらく，全体的にサイズの小さいクラスタが多くなり，クラスタリング終了条件がうまく機能せず全体のクラスタ数が実際のフレーム数より少なくなるためだと考えられる．

4 まとめと今後の展望

本研究では，テキスト中の動詞が喚起する意味フレームの推定において，マスクされた単語の埋め込みと 2 段階クラスタリングを利用する手法を提案した．また，FrameNet を用いた評価実験を行い，従来手法より高い精度を達成すること，マスクされた単語の埋め込みおよび 2 段階クラスタリングが有効であることを示した．

本研究の最終目標はフレーム知識を構築することであり，このためには動詞の意味フレームを推定するだけでなく，各フレームが必要とする項の特定，および，項の意味役割の推定が必要となる．今後は文脈化単語埋め込みをフレーム項の特定，および，その意味役割推定に取り組む予定である．

謝辞

本研究の一部は JSPS 科研費 18H03286 の助成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 *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (ACL-COLING'98)*, pp. 86–90, 1998.

[2] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'18)*, pp. 2227–2237, 2018.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'19)*, pp. 4171–4186, 2019.

[4] Behrang QasemiZadeh, Miriam R. L. Petruck, Regina Stodden, Laura Kallmeyer, and Marie Candito. SemEval-2019 task 2: Unsupervised lexical frame induction. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval'19)*, pp. 16–30, 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 *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval'19)*, 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 *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval'19)*, 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 *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval'19)*, 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] Dan Pelleg, Moore, and Andrew W. X-means: Extending k-means with efficient estimation of the number of clusters. In *Proceedings of the 17th International Conference on Machine Learning (ICML'00)*, pp. 727–734, 2000.

[10] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NIPS'13)*, pp. 3111–3119, 2013.

[11] Chris Biemann. Chinese whispers-an efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of TextGraphs: the First Workshop on Graph Based Methods for Natural Language Processing*, pp. 73–80, 2006.