Vol.2020-NL-244 No.5
2020/7/3

フレーム知識の自動獲得に向けた文脈化単語埋め込みの有用性の検証

山田 康輔[1,a)]　笹野 遼平[1,2]　武田 浩一[1]

概要：本研究では，大規模コーパスからのフレーム知識獲得において，コーパスから収集された動詞の文脈を考慮することの有用性を検証する．具体的には，FrameNet および PropBank において 2 種類以上のフレームを喚起する動詞に着目し，それらの動詞が喚起するフレームの違いを ELMo や BERT に代表される文脈化単語埋め込みがどのくらい捉えているかを，各用例の文脈化単語埋め込みのクラスタリング結果とそれらに付与されたフレームを比較することにより調査する．

キーワード：フレーム知識，文脈化単語埋め込み，FrameNet，PropBank

1. はじめに

自然言語テキストの意味を計算機で扱う際に有用となるリソースの 1 つにフレーム知識があり，FrameNet[*1][1] や PropBank[*2][2] などの大規模なフレーム知識が人手で整備されている．このうち FrameNet では，フレームは特定のイベントや概念を表し，各フレームの情報にはそのフレームが必要とする意味役割等の情報に加えて，そのフレームを喚起しうる語 (フレーム喚起語) の一覧も記載されている．たとえば，動詞「support」は，図 1 に示す Supporting フレームや Evidence フレームのフレーム喚起語である．また，FrameNet では，フレームの情報が人手でタグ付けされた文例も一緒に公開されている．動詞「support」を含む文例を (1)，(2) に示す．文 (1) では「support」は「人やものを支える」という意味で使用されており，喚起するフレームとして Supporting フレームがタグ付けされている．一方，文 (2) では「support」は「証拠となる」という意味で使用されており，Evidence フレームがタグ付けされている．

(1) This study is supported by the fund. (Supporting)

(2) Our results support the hypothesis. (Evidence)

カバレッジの大きいフレーム知識を人手で整備するのは多大なコストがかかることから，このようなフレーム知識を大規模コーパスから自動獲得する取り組みも行われている．たとえば，河原らは大規模コーパスから収集した動詞と項の用例を Chinese Restaurant Process を用いてクラスタリングすることにより述語の意味ごとにフレームを獲得する手法を提案している [3]．また，近年では単語埋め込み表現を用いた手法も提案されており，たとえば Ustalov らは，大規模コーパスから動詞，主語，目的語から成る 3 つ組を収集し，それらの 3 つ組の単語埋め込みを連結した表現を用いてグラフクラスタリングを行うことによりフレーム知識を獲得する手法を提案している [4]．しかし，これらの手法では，まず動詞と項を収集し，収集した動詞や項の文脈を考慮しない単語の意味に基づきクラスタリングを行っており，文脈による動詞や項の意味の違いをほとんど考慮できていない．

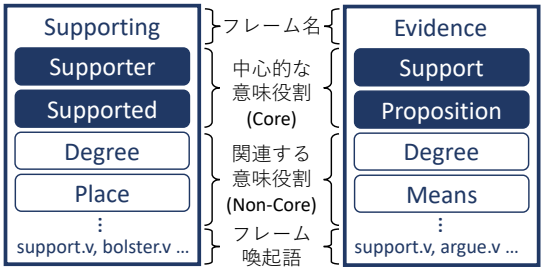

図 1　FrameNet におけるフレームの例．

そこで本研究では，ELMo[5] や BERT[6] などのような文脈化単語埋め込みを利用することで，フレーム知識の自動獲得において文脈を考慮することを考える．文脈化単語埋め込みは周囲の文脈を考慮した表現になっていることから語義曖昧性解消 [7] や異なる文の同一単語に対する用法

[1] 名古屋大学 大学院情報学研究科
[2] 理化学研究所 革新知能統合研究センター
[a)] yamada.kosuke@c.mbox.nagoya-u.ac.jp
[*1] https://framenet.icsi.berkeley.edu/fndrupal/
[*2] https://catalog.ldc.upenn.edu/LDC2013T19

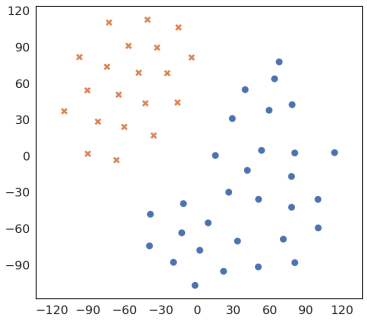

図 2 動詞「support」の BERT による文脈化単語埋め込みを t-SNE により 2 次元にマッピングした結果．●が Supporting フレームを，× が Evidence フレームを喚起する用例を表している．

類似度推定 [8] などのタスクで高い性能を示すことが報告されており，動詞が喚起するフレームの違いの認識にも有用であると考えられる．実際に，FrameNet の文例から動詞「support」を含む文を抽出し，事前学習済みの BERT を用いて，各文における「support」の文脈化単語埋め込みを獲得し，t-SNE[9] により 2 次元にマッピングした結果を図 2 に示す．図 2 から同一のフレームを喚起する「support」の文脈化単語埋め込みが意味ベクトル空間上において固まって分布していることが確認できる．

我々は，文脈化単語埋め込みが持つこのような性質に着目し，文脈化単語埋め込みを利用したフレーム獲得手法の確立を目指す．本研究では，その第 1 ステップとして，ELMo や BERT をはじめとして数多く提案されている文脈化単語埋め込み獲得手法のうち，どの手法がフレームの自動獲得に適しているのか，また，どの程度の精度で動詞の用例によるフレームの違いを認識できるのかを調査する．

2. 関連研究

文脈化単語埋め込みは事前に膨大な数のテキストから言語知識をモデルに学習させることで獲得できる．文脈化単語埋め込みを利用したモデルは，様々な自然言語処理タスクおいて高い精度を実現することが報告されており，多くの文脈化埋め込み獲得手法が提案されている．その先駆けとなったのは ELMo[5] である．ELMo ではまず 2 層の BiLSTM を構築し，その隠れ層の加重平均をとることにより文脈化単語埋め込みを獲得する．近年では，多層の Transformer を用いたモデルが多く提案されている．その代表として広く利用されている BERT[6] では，機械的にマスクした文の一部の予測と次文予測を行うことで文脈化単語埋め込みを獲得する．さらに，BERT を発展させたモデルとして，学習条件を緻密化した RoBERTa[10] やパラメータ数を減らした軽量版である ALBERT[11] が提案されている．他にも，次の単語が何であるかを事前学習させる自己回帰言語モデルに基づく GPT-2[12] や，BERT に対して次文予測を行わず，マスク穴埋めの代わりに自己回帰による学習を行った XLNet[13] などが存在する．

フレーム知識の自動構築において，文脈化単語埋め込みを利用しようとする取り組みはすでにいくつか行われている．特にフレームの自動構築を目的とする SemEval2019 の共通タスク [14] に関連して，ELMo や BERT を利用して動詞横断的なフレーム推定を行った研究が行われている [15][16]．しかし，この共通タスクのデータセットは，同じフレームを喚起する異なる動詞の用例を多く含んでいることから，動詞横断的なフレーム推定の評価には適しているものの，複数のフレームを喚起する動詞に対し，各フレームの用例が十分に含まれているケースは限定的であり，同じ動詞が喚起するフレームの違いを分析するのに適したデータとはなっていない．実際，このタスクに取り組んだ研究では，多くの動詞が 1 つのフレームしか喚起しないという前提を置いており，本研究で着目する動詞の喚起するフレームの違いに関する詳細な分析は為されていない．

本研究で取り組むタスクは，複数のフレームを喚起しうる動詞に対し，動詞の各用例が喚起するフレームの違いを認識するタスクであり，語義曖昧性解消タスクの 1 種と考えることができる．語義曖昧性解消タスクを対象に，BERT による文脈化単語埋め込みの有用性を示した研究は存在しているが [7]，喚起するフレームの同定に焦点を当て，複数の文脈化単語埋め込み手法を比較している点は本研究独自のものである．

3. 文脈化単語埋め込みの有用性の検証

フレーム知識の自動獲得における文脈化単語埋め込みの有用性検証の第 1 ステップとして，動詞の用例によるフレームの違いを文脈化単語埋め込みが捉えているか調査する．具体的には，人手で整備されたフレーム知識リソースにおいて，2 つ以上のフレームを喚起する動詞を対象とし，フレーム情報がタグ付けされた文例集合における文脈化単語埋め込みを獲得し，クラスタリングを行った結果が，タグ付けされたフレームの情報とどの程度一致しているかを調査する．

3.1 フレーム知識リソース

本研究では，フレーム知識として，FrameNet と PropBank を用いる．FrameNet は Fillmore が提唱したフレーム意味論 [17] に基づくフレーム知識リソースであり，FrameNet におけるフレームは複数のフレーム喚起語で共有される．また，関係性が深い一部のフレームの間には「Inheritance」や「Using」などの階層関係が定義されている．一方，PropBank は学習データとして使用できる意味役割タグ付きコーパスの構築を主な目的として開発されたフレーム知識であり，フレームは動詞ごとに固有のものとして定義されている．また，FrameNet のようなフレーム間関係は定義されておらず，それぞれのフレームは独立している．

3.2 検証手順

検証は分析対象とする動詞ごとに以下の手順で行う．

(1) 人手で正しいフレームが付与された文例集合に対し，分析対象の動詞の文脈化単語埋め込みを獲得する[*3]．

(2) 獲得した文脈化単語埋め込みを混合ガウス分布によりクラスタリングする．クラスタ数として利用した文例にタグ付けされたフレームの異なり数を与える．

(3) 人手で付与されたフレームとの一致数が最大になるように各クラスタにフレーム名を重複なく割り当て，そのときのフレーム一致率をスコアとして算出する．

4. 実験

4.1 分析対象とする動詞および文例

本研究では，FrameNet および PropBank からそれぞれ分析対象の動詞，および，文例を抽出し用いる．分析対象とする動詞は，人手で正しいフレームが付与された文例において，2 種類以上のフレームに対し，そのフレームがタグ付けされた文例が 20 件以上含まれている動詞である．たとえば，FrameNet において，動詞「support」は，Supporting フレームがタグ付けされた文例が 30 件，Evidence フレームがタグ付けされた文例が 20 件，存在していることから分析対象とする．一方，動詞「attend」には Attending，Attention，Perception_active と 3 種類のフレームがタグ付けされた文例がそれぞれ存在するが，Attention フレームがタグ付けされた文例は 7 件，Perception_active フレームがタグ付けされた文例は 4 件であり，20 件以上の文例が含まれているフレームは Attending フレームのみであることから分析対象としない．

分析対象の各動詞の文例としては，20 件以上の文例が含まれているフレームの文例のみを使用する．ただし，条件を満たすフレームが 10 を超える場合は文例の多いものから順に 10 フレームの文例のみを使用する．また，フレームごとの文例は最大 100 件とし，100 件を超える文例が存在する場合は無作為に 100 件を選択し用いる．

実際に，FrameNet，および，PropBank から上記の手順により分析対象とする動詞を抜き出した結果，それぞれ 178 個，164 個の動詞が分析対象となった．本研究ではいずれの場合も 120 個の動詞を最終的な評価に使用し，残りをパラメータ調整に使用した．

4.2 実験設定

比較する文脈化単語埋め込みとして，ELMo, $\text{BERT}_{\text{BASE}}$, $\text{BERT}_{\text{LARGE}}$, RoBERTa, ALBERT, GPT-2, および，XLNet を使用した．いずれも公開されている事前学習済みモデルが存在し，ELMo に関しては AllenNLP[*4]内の Original モデルを，その他の Transformer ベースの事前学習済みモデルは Hugging Face[*5]に含まれるモデル[*6]を利用した．埋め込み表現の次元数は $\text{BERT}_{\text{LARGE}}$ のみ 1024 次元，それ以外は 768 次元である．埋め込み表現として使用する隠れ層は，FrameNet，PropBank，それぞれについて調整用データで最も高いスコアを達成した層とした．表 1 にそれぞれのモデルの事前学習に使用されたコーパスのサイズ，学習するパラメータ数，FrameNet，PropBank，それぞれのデータで埋め込み表現として使用した隠れ層番号を示す．

また，混合ガウス分布は scikit-learn の実装[*7]を利用した．この際，Covariance type は spherical と指定した．すなわち，各ガウス分布は次元によらず共通の分散を 1 つだけ持つことになる．

表 1　文脈化単語埋め込みモデルの詳細．埋め込み表現の獲得に使用した隠れ層の番号は左が FrameNet，右が PropBank を表す．

モデル	コーパス	パラメータ	層の数	使用した層
ELMo	11GB	94M	2	2|2
$\text{BERT}_{\text{BASE}}$	16GB	110M	12	10|10
$\text{BERT}_{\text{LARGE}}$	16GB	340M	24	21|17
RoBERTa	160GB	125M	12	9|11
ALBERT	16GB	11M	12	8|10
XLNet	158GB	110M	12	6|5
GPT-2	40GB	117M	12	12|5

表 2　モデルごとの平均スコア．

モデル	FrameNet	PropBank
All-in-one-cluster	0.578	0.548
ELMo	0.627	0.612
$\text{BERT}_{\text{BASE}}$	0.767	0.761
$\text{BERT}_{\text{LARGE}}$	**0.780**	**0.765**
RoBERTa	0.600	0.640
ALBERT	0.714	0.702
GPT-2	0.565	0.545
XLNet	0.718	0.747

4.3 実験結果

表 2 にモデルごとの平均スコアをまとめる．これらのスコアは動詞ごとに算出された一致率の平均をとった値である．All-in-one-cluster は，すべての用例を 1 つのクラスタにまとめた場合のスコアを表し，用例数がもっとも多いフレームが全体に占める割合の平均値となっている．

FrameNet，PropBank，いずれのデータに対しても $\text{BERT}_{\text{LARGE}}$ により文脈化単語埋め込みを獲得した場合がもっとも高いスコアとなった．FrameNet ではそのスコアは 0.78 となっており，$\text{BERT}_{\text{LARGE}}$ による文脈化単語埋め込みは，フレームの自動獲得において有用であると考えられる．$\text{BERT}_{\text{BASE}}$ も $\text{BERT}_{\text{LARGE}}$ とほぼ同じスコア，

[*3] トークン化は事前学習で使用されたのと同様の方法で行い，対象動詞が複数のトークンに分割された場合はその先頭のトークンの文脈化単語埋め込みを使用した．
[*4] https://allennlp.org/elmo
[*5] https://github.com/huggingface/transformers
[*6] それぞれ，bert-base-uncased，bert-large-uncased，roberta-base，albert-base-v2，gpt2，xlnet-base-cased でモデルを指定した．
[*7] https://scikit-learn.org

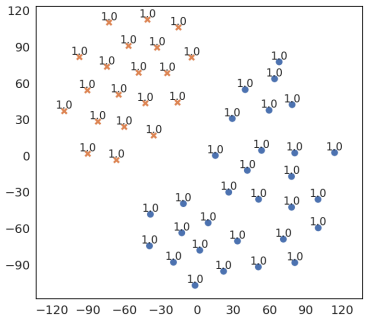

図 3　動詞「support」の用例のクラスタリング結果．●が Supporting フレームを，× が Evidence フレームを喚起する用例を表す．

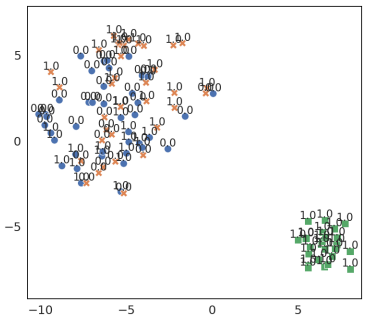

図 4　動詞「fire」の用例のクラスタリング結果．●が Shoot_projectiles フレームを，× が Use_firearm フレームを，■が Firing フレームを喚起する用例を表す．

XLNet，ALBERT がそれらに準ずるスコアとなった一方で，ELMo，RoBERTa，GPT-2 は相対的に低いスコアとなった．この結果から，文脈化単語埋め込みがどのくらいフレームの違いを捉えているかは，埋め込みの獲得手法によって大きく異なることが分かる．どのようなタイプの文脈化単語埋め込みモデルが高いスコアを達成できるかに関する分析は今後の課題である．

図 3，図 4 に，$\text{BERT}_{\text{LARGE}}$ による文脈化単語埋め込みを t-SNE により 2 次元にマッピングし，各点に正解フレームへの負担率を記した結果を示す．負担率は，各用例が正しいフレームに属する割合を表している．図 3 に示した結果は動詞「support」に対する結果である．意味ベクトル空間において Supporing フレームがタグ付けされた用例と Evidence フレームがタグ付けされた用例が，それぞれ固まって分布しており，タグ付けされたフレームごとに適切にクラスタが構築されていることが確認できる．一方，図 4 に示した結果は，動詞「fire」に対する結果であるが，Firing フレームがタグ付けされた用例は 1 つのクラスタを構成しているのに対し，Shoot_projectiles フレーム，Use_firearm フレームの違いを捉えられていないことが確認できる．ここで，Firing フレームは「解雇する」という意味を表すフレームであり，他の 2 つのフレームの表す内容と大きく異なっている．一方，Shoot_projectiles フレームと Use_firearm はそれぞれ「弾丸を撃つ」,「銃を撃つ」を表すフレームであり，前者と後者の間には「Using」というフレーム間関係が定義されている非常に関係の深いフレームとなっており，文脈からこれらのフレームの違いを識別するのは非常に難しいと考えられる．

4.4　フレーム間関係の有無による精度の関係

動詞「fire」の例からも明らかなように，フレーム間関係が定義されたフレームは，その文脈が類似しており，文脈化単語埋め込みを利用してもそれらの用法を区別するのは難しい場合が多いと考えられる．加えて，これらのフレームを別のフレームとして定義するべきかどうかにはある程度の恣意性が存在していると考えられ，フレームの自動獲得を考えた場合，これらのフレームを識別する重要性は，Firing フレームと Use_firearm フレームのようにまったく別の意味を表すフレームより小さいと考えられる．そこで，FrameNet から抽出した動詞を対象に，フレーム間関係が定義されているかどうかが，用例のクラスタリング精度にどのくらい影響するかを調査した．

表 3　FrameNet におけるフレーム間関係の有無と平均スコア．

モデル	関係無	関係有	差分
All-in-one-cluster	0.604	0.589	0.015
ELMo	0.655	0.634	0.021
$\text{BERT}_{\text{BASE}}$	0.801	0.725	0.076
$\text{BERT}_{\text{LARGE}}$	0.801	0.764	0.037
RoBERTa	0.611	0.636	-0.025
ALBERT	0.753	0.657	0.096
GPT-2	0.589	0.603	-0.014
XLNet	0.758	0.692	0.066

具体的には，実験に用いた動詞のうち，用例にタグ付けされたフレームの種類数がちょうど 2 つである動詞を対象に，それら 2 つのフレームの間にフレーム間関係がある動詞とない動詞で分けてスコアを算出した．ここで，フレーム数が 2 つである動詞に対象を限定したのは，フレーム数が多いほどスコアが低くなりやすいというタスクの性質の影響を無視するためである．結果を表 3 に示す．4.3 節の表 2 に示した結果において相対的に高い精度を達成した $\text{BERT}_{\text{LARGE}}$，$\text{BERT}_{\text{BASE}}$，ALBERT，XLNet の 4 つのモデルにおいては，フレーム間関係がないフレームの識別の方が高い精度となっていることが確認できる．より重要性の高いフレームの識別に成功しているということからも，これら 4 種の文脈化単語埋め込みはフレームの違いを適切に捉えていると考えられる．

5.　フレーム数の自動推定

前節までの実験では，正しいフレーム数を人手で与えていたが，実際にフレームの自動獲得を行う場合はフレーム数も自動推定することが必要となる．そこで，クラスタ数の推定に広く用いられているベイズ情報量規準 (BIC)[18],

および，出力されるフレーム数が FrameNet，PropBank でタグ付けされたフレーム数に近くなるように BIC を調整した調整版 BIC (ajusted-BIC) を用いて，フレーム数を自動推定できるか調査する．

5.1 フレーム数の推定方法

教師無しクラスタリングにおいてフレーム数の自動推定に用いられる尺度にベイズ情報量規準 (BIC) がある．BIC は次の式で定義される．

$$\mathrm{BIC} = -2\ln(L) + k \cdot \ln(n_s) \tag{1}$$

ここで，L はモデルの尤度，n_s は用例数，k はモデルのパラメータ数を表す．本研究で使用している混合ガウス分布では Covariance type として spherical を使用していることから，ガウス分布 1 つあたりの分散を表すパラメータ数は 1 つである．また，混合ガウス分布は各ガウス分布の中心点を表すために埋め込み次元数 d のパラメータを持ち，さらに，総和が 1 となるような各ガウス分布への重み付けパラメータを持つ．したがって，クラスタ数を n_c とした場合，全体のパラメータ数は $k = (d+2) \times n_c - 1$ となる．BIC をクラスタ数の決定に使用する場合，一般的に BIC の値がもっとも小さくなるクラスタ数に決定する．一般的に最適化されたモデルの尤度はパラメータ数が多くなるにつれ大きくなることから，BIC の右辺の第 1 項はクラスタ数の増加に伴い減少する傾向にある．したがって，右辺の第 2 項はクラスタ数が増えることを抑制するペナルティ項であるとみることができる．

フレームの粒度は人の直観に基づき設定されたものであり，必ずしも情報量規準で最適とされる粒度となっているとは限らないと考えられる．そこで本研究ではフレームの粒度が与えられたデータに近くなるように BIC のペナルティ項を調整した調整版 BIC (ajusted-BIC) を導入する．調整版 BIC は下記の式で与えられる．

$$\text{ajusted-BIC} = -2\ln(L) + c \cdot k \cdot \ln(n_s) \tag{2}$$

この式で用いられている c はペナルティ項の大きさを調整するパラメータであり，調整用データを利用して決定する．

5.2 フレーム数の自動推定実験

前節までの実験では，フレーム数が 2 つ以上存在する動詞のみを対象としていたが，フレーム数推定においては，フレームを 1 つしか持たないことを正しく認識することも重要であることから，フレーム数が 1 つである動詞も追加する．具体的には，用例を 20 件以上含むフレームを 1 つしか持たない動詞の中で用例数が多いものを，前節までの実験で使用した数と同数追加する．使用する用例数は前節までと同様に上限 100 件とした．したがって，FrameNet を用いた実験ではパラメータ調整に 116 個，評価に 240 個の動詞を，PropBank を用いた実験ではパラメータ調整に 88 個，評価に 240 個の動詞を用いることになる．

表 4 FrameNet および PropBank に対して BIC，調整版 BIC (a-BIC) を用いた場合のフレーム数の推定精度．

	FrameNet		PropBank	
	BIC	a-BIC	BIC	a-BIC
正解率	0.063	0.513	0.004	0.604
RMSE	6.429	2.026	5.733	1.340
ρ	0.104	0.241	0.332	0.499

表 5 FrameNet データに対する混同行列．縦軸が人手で付与されたフレーム数，横軸が推定されたクラス数を表す．

	1	2	3	4	5≤
1	74	32	6	1	7
2	40	48	5	0	6
3	3	14	1	0	0
4	0	1	2	0	0

表 6 PropBank データに対する混同行列．縦軸が人手で付与されたフレーム数，横軸が推定されたクラス数を表す．

	1	2	3	4	5≤
1	90	24	2	0	4
2	29	46	7	1	0
3	3	3	8	0	1
4 ≤	3	10	4	1	4

文脈化単語埋め込みには $\mathrm{BERT}_{\mathrm{LARGE}}$ を使用した．また，フレーム数の自動推定の評価は，各用例に人手で付与されたフレームの異なり数との一致率 (正解率)，二乗平均平方根誤差 (RMSE)，スピアマンの順位相関係数 (ρ) を用いて行った．フレームの数の自動推定結果を表 4 に示す．FrameNet，PropBank，いずれのデータに対しても，ペナルティ項の大きさを調整した調整版 BIC を用いた場合の方が高いスコアとなった．このとき，実際に調整されたパラメータ c の値は FrameNet で 2.4，PropBank で 3.3 であり，パラメータを導入しない場合はクラスタを多く生成する傾向にあった．

FrameNet，PropBank，それぞれについて人手で付与されたフレーム数と，調整版 BIC に基づき自動推定されたフレーム数の混同行列を表 5，表 6 に示す．FrameNet，PropBank に対する順位相関係数はそれぞれ 0.241，0.499 と弱い相関は確認できる．しかし，人手で付与されたフレームが 1 つのみである用例集合に対し，5 つ以上のクラスタを生成するケースも存在しており十分に高い精度とは言い難いのが現状であるが，今後，動詞横断的なフレーム獲得を行うことで，自然に改善される可能性も考えられる．

6. おわりに

本研究では，フレーム知識の自動獲得に向けて，文脈化単語埋め込みの有用性の検証を行った．具体的には，フレーム知識リソースにおいて 2 つ以上のフレームを喚起する動

詞を対象として，それらの用例を文脈化単語埋め込みに基づきクラスタリングし，構築されたクラスタとフレームの一致率を評価した．実験は FrameNet と PropBank の 2 つのフレーム知識リソースを対象に，7 種類の文脈化単語埋め込みを用いて行い，BERT や XLNet，ALBERT による文脈化単語埋め込みが，人手で付与されたフレームの違いを高い精度で捉えていることを示した．またベイズ情報量規準を調整した指標を用いることでフレーム数の自動推定を 50%〜60%の精度で行うことができることを確認した．

本研究では，同一の動詞が異なるフレームを喚起するケースに着目し，フレームの自動獲得における文脈化単語埋め込みの有用性の検証を行ったが，FrameNet のような動詞横断的なフレーム知識を構築する場合は，動詞横断的な調査も必要となる．また，フレーム知識を獲得するためには，動詞をその動詞が喚起するフレームごとに分けるだけではなく，項の意味役割をまとめる必要がある．項の文脈化単語埋め込みは，その意味役割を捉えていると考えられ，クラスタリングを行うことで共通の意味役割をまとめることができると考えられる．さらに，本研究では，動詞のみを対象として検証を行っているが，フレームを喚起する名詞や形容詞など別の品詞に対しても同様の結果を得られるか調査を行う必要がある．これらの調査を行うことで，目標とする高品質なフレーム知識の自動獲得に近づくことが期待できる．

参考文献

[1] Baker, C. F., Fillmore, C. J. and Lowe, J. B.: The Berkeley FrameNet Project, *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] Kingsbury, P. and Palmer, M.: From TreeBank to PropBank, *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, pp. 1989–1993 (2002).

[3] Kawahara, D., Peterson, D., Popescu, O. and Palmer, M.: Inducing Example-based Semantic Frames from a Massive Amount of Verb Uses, *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL'14)*, pp. 58–67 (2014).

[4] Ustalov, D., Panchenko, A., Kutuzov, A., Biemann, C. and Ponzetto, S. P.: Unsupervised Semantic Frame Induction using Triclustering, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL'18)*, pp. 55–62 (2018).

[5] Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L.: Deep Contextualized Word Representations, *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).

[6] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *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).

[7] Hadiwinoto, C., Ng, H. T. and Gan, W. C.: Improved Word Sense Disambiguation Using Pre-Trained Contextualized Word Representations, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'19)*, pp. 5297–5306 (2019).

[8] Garí Soler, A., Apidianaki, M. and Allauzen, A.: Word Usage Similarity Estimation with Sentence Representations and Automatic Substitutes, *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM'19)*, pp. 9–21 (2019).

[9] Maaten, L. v. d. and Hinton, G.: Visualizing Data using t-SNE, *Journal of Machine Learning Research*, Vol. 9, pp. 2579–2605 (2008).

[10] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V.: RoBERTa: A Robustly Optimized BERT Pretraining Approach, arXiv preprint arXiv:1907.11692 (2019).

[11] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P. and Soricut, R.: ALBERT: A Lite BERT for Self-supervised Learning of Language Representations, arXiv preprint arXiv:1909.11942 (2019).

[12] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I.: Language Models are Unsupervised Multitask Learners, *OpenAI Blog*, Vol. 1, No. 8 (2019).

[13] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R. and Le, Q. V.: XLNet: Generalized Autoregressive Pretraining for Language Understanding, *Advances in Neural Information Processing Systems (NIPS'19)*, pp. 5754–5764 (2019).

[14] QasemiZadeh, B., Petruck, M. R. L., Stodden, R., Kallmeyer, L. and Candito, M.: SemEval-2019 Task 2: Unsupervised Lexical Frame Induction, *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval'19)*, pp. 16–30 (2019).

[15] Anwar, S., Ustalov, D., Arefyev, N., Ponzetto, S. P., Biemann, C. and Panchenko, A.: HHMM at SemEval-2019 Task 2: Unsupervised Frame Induction using Contextualized Word Embeddings, *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval'19)*, pp. 125–129 (2019).

[16] Arefyev, N., Sheludko, B., Davletov, A., Kharchev, D., Nevidomsky, A. and Panchenko, A.: Neural GRANNy at SemEval-2019 Task 2: A Combined Approach for Better Modeling of Semantic Relationships in Semantic Frame Induction, *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval'19)*, pp. 31–38 (2019).

[17] Fillmore, C. J.: Frame Semantics, *Cognitive Linguistics: Basic Readings*, Vol. 34, pp. 373–400 (2006).

[18] Schwarz, G. et al.: Estimating the Dimension of a Model, *The Annals of Statistics*, Vol. 6, No. 2, pp. 461–464 (1978).