Verb Sense Clustering using Contextualized Word Representations for Semantic Frame Induction

Kosuke Yamada[1] **Ryohei Sasano**[1,2] **Koichi Takeda**[1]
[1]Graduate School of Informatics, Nagoya University, Japan
[2]RIKEN Center for Advanced Intelligence Project, Japan
yamada.kosuke@c.mbox.nagoya-u.ac.jp,
{sasano,takedasu}@i.nagoya-u.ac.jp

Abstract

Contextualized word representations have proven useful for various natural language processing tasks. However, it remains unclear to what extent these representations can cover hand-coded semantic information such as semantic frames, which specify the semantic role of the arguments associated with a predicate. In this paper, we focus on verbs that evoke different frames depending on the context, and we investigate how well contextualized word representations can recognize the difference of frames that the same verb evokes. We also explore which types of representation are suitable for semantic frame induction. In our experiments, we compare seven different contextualized word representations for two English frame-semantic resources, FrameNet and PropBank. We demonstrate that several contextualized word representations, especially BERT and its variants, are considerably informative for semantic frame induction. Furthermore, we examine the extent to which the contextualized representation of a verb can estimate the number of frames that the verb can evoke.

1 Introduction

Contextualized word representations such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) are known to be effective in many natural language processing tasks such as question answering, natural language inference, and semantic textual similarity. Contextualized word representations can generate different representations of the same word in different contexts and distinguish the polysemy of a word. It has been reported that this property is effective in word sense disambiguation (WSD) (Hadiwinoto et al., 2019) and word sense induction (WSI) (Amrami and Goldberg, 2018). Therefore, it appears that contextualized word representations can also be leveraged to induce semantic frames from a large corpus automatically.

A semantic frame is defined on the basis of the semantic roles that a predicate can take as its arguments. FrameNet[1] (Baker et al., 1998) and PropBank[2] (Palmer et al., 2005) are the two most well-known resources of semantic frames, both of which are manually compiled. These resources are used not only for semantic parsing (Yang and Mitchell, 2017) but also for information extraction (Gangemi et al., 2016), question answering (Shen and Lapata, 2007), and document summarization (Cheung and Penn, 2013).

These frame-semantic resources define the frame and semantic roles, and they provide example sentences in which they are annotated. For example, the verb "support" in FrameNet is defined to evoke two frames: the SUPPORTING frame and the EVIDENCE frame. Sentences (1) and (2) below are examples where these frames are annotated. In Sentence (1), "support" means 'supporting a person or a thing' and evokes the SUPPORTING frame. Its arguments are annotated with the semantic roles of Supporter and Supported. In Sentence (2), "support" means 'corroborating' and evokes the EVIDENCE frame. Its arguments are annotated with the semantic roles of Proposition and Support. In both examples, the frame-evoking word is "support," but its evoking frames are different.

(1) [Supported This study] is supported by [Supporter the fund]. (SUPPORTING)

(2) [Support Our results] support [Proposition the hypothesis]. (EVIDENCE)

Since the manual development of such broad-coverage frame-semantic resources is labor-intensive and time-consuming, many researchers have attempted to induce semantic frames from

[1]https://framenet.icsi.berkeley.edu/
[2]https://propbank.github.io/

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4353–4362

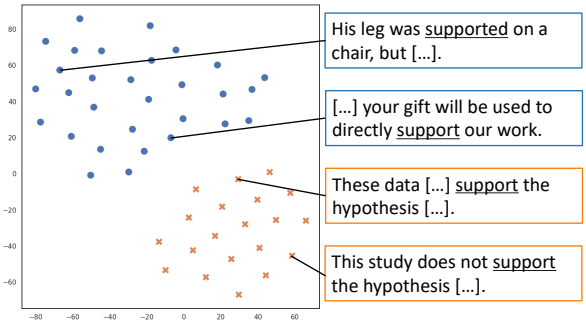

Figure 1: 2D t-SNE projection of BERT representations of verb "support" in FrameNet. ● and × correspond to example sentences from SUPPORTING and EVIDENCE frames, respectively.

large corpora automatically. For example, Kawahara et al. (2014) extracted predicate-argument structures of each verb from large corpora and induced the frames that each verb evokes by clustering the extracted predicate-argument structures. Several researchers have recently proposed frame induction methods that leverage word vector representations. For example, Ustalov et al. (2018) collected subject-verb-object triples from a Web-scale corpus and induced the frames by clustering based on the concatenation of word vector representations of the triples. However, since these approaches first collect the tuples of a verb and its arguments and then perform the clustering based on their word representations without taking their contexts into account, they may fail to disambiguate the word senses that require contextual clues.

Therefore, we seek a frame induction method that makes better use of contextual information by leveraging contextualized word representations. Figure 1 shows a 2D projection of contextualized representations of the verb "support" in different sentences. We extracted example sentences of "support" from the frame-annotated sentences in FrameNet, acquired contextualized representations of the verbs by applying a pre-trained BERT, and then projected them into two dimensions by using t-distributed stochastic neighbor embedding (t-SNE) (Maaten and Hinton, 2008). As shown in the figure, these BERT representations are distributed separately depending on the frame that "support" evokes in each example.

Our objective is to exploit this property of contextualized word representations for semantic frame induction. As a first step, we investigate how well contextualized word representations can distinguish the frames that the same verb evokes and which type of representations are suitable for semantic frame induction. We also need to estimate the number of frames that a verb evokes to build a frame-semantic resource automatically. We clarify to what extent contextualized word representations of verbs can estimate the number of frames that verbs evoke, which are defined manually. Our investigation of contextualized word representations will help construct high-quality frame-semantic resources not only for high-resource languages and general domains but also for low-resource languages and specific domains.

2 Related Work

2.1 Contextualized Word Representations

Contextualized word representations encode semantic and syntactic information by learning linguistic patterns and constraints from a large amount of text and provide significant improvements to the state of the art for a wide range of natural language processing tasks. They are also widely applied as context-sensitive word representation extractors for summarization (Liu, 2019), neural machine translation (Zhu et al., 2019), and so on.

Recently, several contextualized word representations have been proposed. For example, ELMo (Peters et al., 2018) produces contextualized word representations by pre-training on a bidirectional language model task in 2-layer BiLSTMs. More recently, many Transformer-based (Vaswani et al., 2017) models have been proposed. BERT (Devlin et al., 2019) utilizes multilayer bidirectional Transformers and is pre-trained on two tasks: masked language modeling and next sentence prediction. RoBERTa (Liu et al., 2019) redesigns the pre-training conditions for BERT, and ALBERT (Lan et al., 2020) shares each layer's parameters in BERT to reduce the number of parameters. There are other models such as GPT-2 (Radford et al., 2019), which is a unidirectional model that is trained to predict the next word in a sentence, and XLNet (Yang et al., 2019), which is based on a permutation language model that learns a bidirectional context in an autoregressive manner.

2.2 Semantic Frame Induction

For semantic frame induction of a word in a context, it is a standard approach to extract predicate-argument structures and then perform the clustering of those structures. LDA-frames (Materna, 2012) is an approach that represents frames as tuples of

subject and object and uses latent Dirichlet allocation (LDA) (Blei et al., 2003) to induce semantic frames. Kawahara et al. (2014) extracted predicate-arguments structures from a large Web corpus and then applied the Chinese restaurant process clustering-algorithm (Aldous, 1985) to group predicates with similar arguments. Ustalov et al. (2018) proposed the Triclustering, which produces subject-verb-object triples and then performs a graph-based clustering using the concatenations of their static word embeddings. These methods take only the predicates and their arguments into account, and they do not sufficiently consider the context.

In some works, contextualized word representations are already used for semantic frame induction. In a shared task at SemEval 2019 (QasemiZadeh et al., 2019), some researchers worked on an unsupervised semantic frame induction task, and they reported that ELMo and BERT were useful for the task. Arefyev et al. (2019) first performed group average clustering by using contextualized word embeddings of target verbs from BERT. Then, they performed clustering to split each cluster into two by using TF-IDF features with paraphrased words by using BERT. Anwar et al. (2019) used a concatenated representation of a target verb and the average word embedding of all words in a sentence obtained by skip-gram (Mikolov et al., 2013) or ELMo. They performed group average clustering based on Manhattan distance by using the embedding. Ribeiro et al. (2019) performed graph clustering based on Chinese whispers (Biemann, 2006) by using contextualized representations of frame-evoking verbs from ELMo or BERT.

The shared task dataset contains many example sentences in which different verbs evoke the same frame, and thus the dataset is suitable for evaluating semantic frame induction over verbs. However, there are few example sentences of verbs that evoke different frames in the dataset, and it is not ideal for analyzing the difference of frames that each verb evokes. Some researchers assumed that many verbs evoke only one frame, and they did not analyze the difference of frames that each verb evokes.

Also, there is a study that works on semantic frame induction by using contextualized word representations in semi-supervised learning. Yong and Timponi Torrent (2020) used ELMo or BERT and mapped high-dimensional representations of verbs to a low-dimensional latent space for better frame prediction. Their study aims to extend FrameNet. On the other hand, our goal is to build frame-semantic resources automatically in an unsupervised fashion.

2.3 Word Sense Disambiguation with Contextualized Word Representation

The task in this paper is to distinguish the difference of frames that the same verb evokes, and as such, can be regarded as a type of word sense disambiguation (WSD) task. For the WSD task, contextualized word representations have been reported to be useful. For example, Peters et al. (2018) performed the task by nearest neighbor matching with ELMo representations, and Hadiwinoto et al. (2019) used pre-trained BERT contextualized representations as features for WSD. While WSD aims to distinguish between the meanings of words on the same surface, the semantic frame induction we focus on aims to distinguish between intuitive concepts such as situations, objects, and events that words evoke.

3 Methodology

We investigate to what extent contextualized word representations recognize the difference of frames that the same verb evokes. Specifically, we focus on verbs that evoke more than one frame in frame-semantic resources and acquire contextualized word representations of them. We then apply clustering and evaluate how well the generated clusters and human-annotated frames match.

3.1 Frame-semantic Resources

We use FrameNet and PropBank in English as frame-semantic resources. Since our goal is to establish a semantic frame induction method that is not in a particular style, we use two well-known frame knowledge resources for our investigation: Berkeley FrameNet data release 1.7[3] and PropBank-annotated data from OntoNotes v5.0.[4]

FrameNet is developed within the framework of the theory of frame semantics proposed by Fillmore (2006). Each frame is shared by multiple frame-evoking words (lexical units), and hierarchical relations such as "Inheritance" or "Using" are defined between closely related frames. FrameNet has 1,222 frames, 13,572 lexical units, and 200,751 annotated sentences. The corpus consists of the British National Corpus and U.S. newswire texts.

[3] https://framenet.icsi.berkeley.edu/fndrupal/framenet_data

[4] https://catalog.ldc.upenn.edu/LDC2013T19

PropBank is developed as a corpus with semantic roles that can be used as training data in supervised learning. PropBank frames are defined for each verb as a frameset containing semantic role labels. There are two types of labels; one is ARG0-5. It indicates a necessary role and has a different meaning in each frameset. The other is an argument modifier (AM) label, which indicates an additional role common to all framesets (e.g., AM-TMP for time). For example, the frameset SUPPORT.01 (lend aid, credence to) of "support" is defined with ARG0 as 'helper' and ARG1 as 'person, thing or project being supported.' Sentence (3) is an example in which this frameset is annotated.

(3) [$_{\text{ARG1}}$ Students] were supported$_{.01}$ by [$_{\text{ARG0}}$ the scholarship] [$_{\text{AM-TMP}}$ for four years].

Unlike FrameNet, hierarchical relations are not defined between framesets; that is, each frameset is independent. PropBank has 5,607 framesets, 4,221 verbs, and 111,178 annotated sentences. The corpus consists of newswires, magazine articles, broadcast news, broadcast conversations, web data, conversational speech data, and pivot text.

3.2 Procedure

In our investigation, we follow the procedures below for each target verb that evokes more than one frame in the frame-semantic resources.

1. Acquire contextualized word representations of the target verbs in the set of frame-annotated example sentences in the frame-semantic resources.[5]

2. Apply clustering to their contextualized word representations by using a Gaussian mixture model. At this time, the number of clusters given to the model is equal to the number of frames in our dataset.

3. Find a mapping between the generated clusters and the human-annotated frames that maximize the overall number of matches. We use the match rate as the evaluation metrics.

4 Experiment on Frame Distinction

4.1 Dataset

We first determined the target verbs, and we then extracted example sentences of the target verbs from both FrameNet and PropBank. As target verbs, we used verbs that evoke two or more frames with at least 20 annotated sentences. For example, in FrameNet, the verb "support" is a target verb because there are 30 sentences in the SUPPORTING frame and 20 sentences in the EVIDENCE frame. In contrast, the verb "attend" is not a target verb. This is because although the verb "attend" evokes three frames, (ATTENTION, PERCEPTION_ACTIVE, and ATTENDING), there are 7 sentences in the ATTENTION frame, 4 sentences in the PERCEPTION_ACTIVE frame, and 24 sentences in the ATTENDING frame and only the ATTENDING frame includes 20 or more sentences.

For each verb, we considered frames that include at least 20 annotated sentences. In addition, if the target verb evokes more than 10 frames with 20 or more annotated sentences, we used the top 10 frames on the basis of the number of annotated sentences. We used a maximum of 100 annotated sentences for each frame. As a result, we have obtained 178 target verbs for FrameNet and 164 for PropBank. The average counts of frames per verb were 2.21 for FrameNet and 2.73 for PropBank, and the average counts of annotated sentences per frame were 41.68 for FrameNet and 70.34 for PropBank. In this paper, we used 120 verbs as the test set for the final evaluation and the remaining verbs as the development set for tuning the parameters for both FrameNet and PropBank.

4.2 Settings

We compared ELMo, BERT$_{\text{BASE}}$, BERT$_{\text{LARGE}}$, RoBERTa, ALBERT, GPT-2, and XLNet as contextualized word representations in order to explore the representation most suitable for semantic frame induction. We used publicly available pre-trained models. ELMo is the 'Original' model in AllenNLP,[6] and the other transformer-based models are pre-trained models[7] in Hugging Face.[8] For each model, we obtained contextualized word representations from the hidden layer that achieved the highest scores in the development sets for FrameNet and PropBank, respectively. Table 1 lists the size of the corpus used to pre-train models and the number of parameters, dimensions, layers, and hidden layers of models used to obtain the representations

[5] Tokenization is performed in the same way as used in the pre-training. If tokenization splits the target verb token into more than one sub-token, we use the contextualized word representations of the first sub-token.

[6] https://allennlp.org/elmo

[7] These models are specified by 'bert-base-uncased,' 'bert-large-uncased,' 'roberta-base,' 'albert-base-v2,' 'gpt2,' and 'xlnet-base-cased' in Hugging Face.

[8] https://huggingface.co/transformers/

Model	Corpus size	No. of parameters	No. of dimensions	No. of layers	Used layer
ELMo	11GB	94M	512	2	1|1
$\text{BERT}_{\text{BASE}}$	16GB	110M	768	12	9|7
$\text{BERT}_{\text{LARGE}}$	16GB	340M	1024	24	21|15
RoBERTa	160GB	125M	768	12	10|6
ALBERT	16GB	11M	768	12	9|8
GPT-2	40GB	117M	768	12	8|9
XLNet	158GB	110M	768	12	5|5

Table 1: Details of contextualized word representations. "Used layer" means hidden layer of model used to obtain the representations in FrameNet (left) and PropBank (right).

Model	FrameNet	PropBank
All-in-one-cluster	0.578	0.548
ELMo	0.631	0.607
$\text{BERT}_{\text{BASE}}$	0.750	0.765
$\text{BERT}_{\text{LARGE}}$	**0.769**	0.790
RoBERTa	0.767	**0.796**
ALBERT	0.705	0.712
GPT-2	0.666	0.650
XLNet	0.729	0.758

Table 2: Macro-average match rate of each verb for each of models and datasets.

for FrameNet and PropBank, respectively.

We used the Gaussian mixture model implementation provided by scikit-learn.[9] We adopted "spherical" as the covariance type, that is, the covariance matrix was a diagonal covariance with equal elements along the diagonal. We used five trials of clustering with different random seeds and adopted the result of the highest likelihood trial.

4.3 Results

Table 2 lists the macro-average match rate of each verb for each of the models and datasets. All-in-one-cluster means the average score when all the example sentences were in one cluster for each verb. That is, the score is the average of the percentages of examples that were annotated with the most frequently used frame for a verb. For example, the score of the verb "support" in FrameNet was 0.6 (30/50) since the numbers of example sentences from the SUPPORTING frame and the EVIDENCE frame were 30 and 20, respectively.

As shown in Table 2, $\text{BERT}_{\text{LARGE}}$ and RoBERTa achieved the highest scores for FrameNet and PropBank, respectively. We confirmed that they recognized the differences of frames that the same verbs evoke. $\text{BERT}_{\text{BASE}}$, XLNet, and ALBERT also achieved high scores. These results indicate that BERT, RoBERTa, XLNet, and ALBERT are useful for semantic frame induction. In contrast, the scores obtained for ELMo and GPT-2 were relatively low and almost the same as for the All-in-one-cluster. It indicates that the degree of the difference of frames captured by the contextualized word representations varied greatly.

The reasons for these results are described below. The high scoring BERT, RoBERTa, XLNet, and ALBERT are deep bidirectional language models based on Transformer. In contrast, GPT-2 is a unidirectional language model based on Transformer. Also, ELMo is a relatively sparse bidirectional language model that consists of only two unidirectional contexts concatenated together. Therefore, the scores of GPT-2 and ELMo were lower than those of the deep bidirectional language models.

We show several examples below. In these figures, the number given to each point represents the clustering result; that is, the points with the same number belong to the same cluster. Note that the value of the number has no meaning. Figure 2 shows a 2D t-SNE projection of $\text{BERT}_{\text{LARGE}}$ vectors for "support" in FrameNet. We can see that the example sentences from the SUPPORTING frame and the EVIDENCE frame form a cluster, respectively.

Figure 3 shows a 2D t-SNE projection of $\text{BERT}_{\text{LARGE}}$ vectors for "fire" in FrameNet. We can see that example sentences from the FIRING frame form a single cluster, whereas the difference between the SHOOT_PROJECTILES frame and the USE_FIREARM frame is not captured. The FIRING frame means 'ending an employment relationship' while the SHOOT_PROJECTILES and the USE_FIREARM frames mean 'shooting a bullet' and 'shooting a gun', respectively. The FIRING frame is very different from the other two. On the other hand, the "Using" relation is annotated between the SHOOT_PROJECTILES frame and USE_FIREARM frame, which indicates that there is a strong connection between the two frames. We conduct an additional analysis on frames that have hierarchical

[9] https://scikit-learn.org

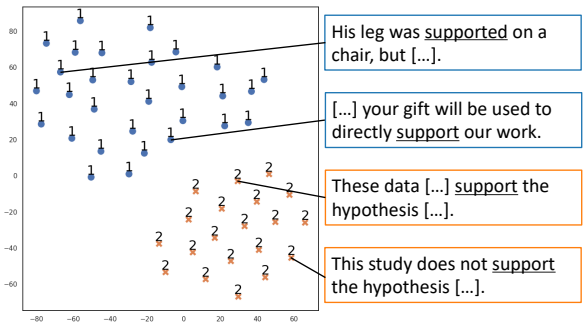

Figure 2: 2D t-SNE projection of $\text{BERT}_{\text{LARGE}}$ vectors for verb "support" in FrameNet. ● and × correspond to example sentences from SUPPORTING and EVIDENCE frames, respectively.

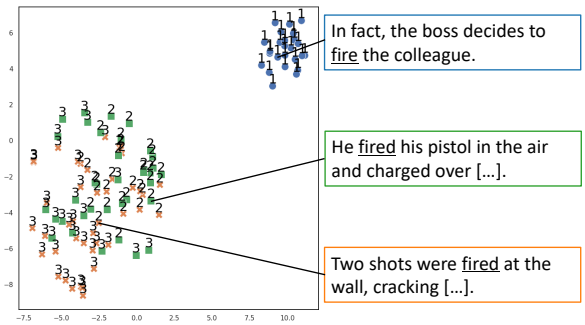

Figure 3: 2D t-SNE projection of $\text{BERT}_{\text{LARGE}}$ vectors for verb "fire" in FrameNet. ●, ×, and ■ correspond to example sentences from FIRING, SHOOT_PROJECTILES, and USE_FIREARM frames, respectively.

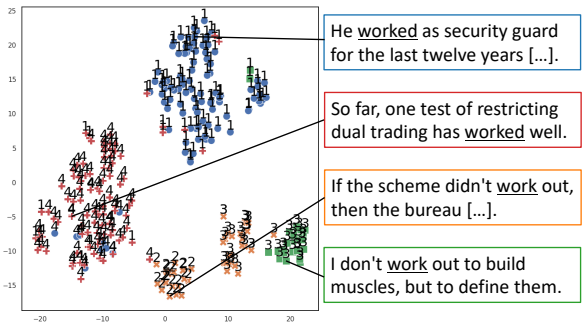

Figure 4: 2D t-SNE projection of $\text{BERT}_{\text{LARGE}}$ vectors for verb "work" in PropBank. ●, ×, ■, and + correspond to example sentences from WORK.01 (work), WORK.02 (arrange), WORK.03 (exercise), and WORK.09 (function, operate) framesets, respectively.

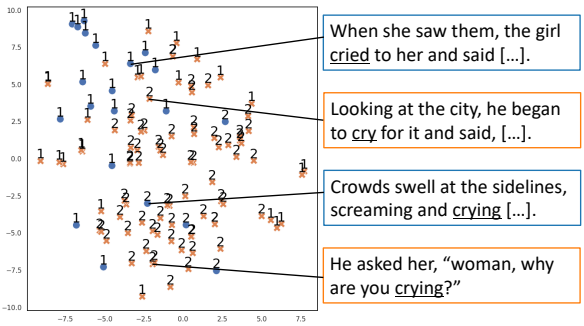

Figure 5: 2D t-SNE projection of $\text{BERT}_{\text{LARGE}}$ vectors for verb "cry" in PropBank. ● and × correspond to example sentences from CRY.01 (speak loudly, yell, demand, possibly while weeping) and CRY.02 (cry, weep) framesets, respectively.

relations in Section 4.4.

Figure 4 shows a 2D t-SNE projection of $\text{BERT}_{\text{LARGE}}$ vectors for "work" in PropBank. The verb "work" has four types of framesets: WORK.01 (work), WORK.02 (arrange), WORK.03 (exercise), and WORK.09 (function, operate). We confirmed that $\text{BERT}_{\text{LARGE}}$ roughly captured the difference of frames, even for verbs that can have many framesets. In the examples where WORK.02 and WORK.03 were annotated, the verb "work" appears in the form of "work out," and it may have been a bit challenging to capture the difference of these framesets. This is because verbs that appear as part of phrasal verbs have relatively similar contextualized word embeddings since the same word appears near the verb.

Figure 5 shows a 2D t-SNE projection of $\text{BERT}_{\text{LARGE}}$ vectors for "cry" in PropBank. The verb "cry" has two types of framesets: CRY.01 (speak loudly, yell, demand, possibly while weeping) and CRY.02 (cry, weep). Like the verb "fire" in FrameNet, the resulting clusters could not be appropriately formed because the framesets of the verb "cry" are both related to 'weep' and are thus very similar.

4.4 Effect of Hierarchical Relations on Evaluation

The frames with hierarchical relations defined in FrameNet appear in similar contexts. As is clear from the examples of "fire," it is not easy to distinguish these frames, even using contextual word representations. Moreover, it is unclear whether these frames should be defined as separate frames if semantic frame resources are to be automatically constructed in the future. Specifically, the importance of distinguishing between the SHOOT_PROJECTILES frame and the USE_FIREARM frame could be less important than distinguishing between the SHOOT_PROJECTILES frame and the FIRING frame.

To investigate the practical usefulness, we attempted to evaluate the accuracy of the distinc-

Model	Gr w/o rel	Gr w/ rel	Diff
All-in-one-cluster	0.613	0.604	0.009
ELMo	0.680	0.644	0.036
$\text{BERT}_{\text{BASE}}$	0.805	0.719	0.086
$\text{BERT}_{\text{LARGE}}$	0.826	0.713	0.113
RoBERTa	0.836	0.707	0.129
ALBERT	0.758	0.670	0.088
GPT-2	0.724	0.655	0.069
XLNet	0.783	0.700	0.083

Table 3: Average match rate by groups without hierarchical relations (Gr w/o rel) and groups with hierarchical relations (Gr w/ rel). "Diff" represents difference between score of Gr w/o rel and score of Gr w/ rel.

tion between frames with hierarchical relations and frames without relations, separately. We first extracted verbs that had exactly two types of frames from FrameNet as a result of the procedure described in Section 4.1. We then divided the extracted verbs into two groups according to whether there is a hierarchical relation between the two frames or not and calculated the average match rate for each group.[10] By limiting our focus to verbs with two types of frames, we can ignore the tendency of the match rate to decrease as the number of frames increases. We assume that if a certain contextualized word representation appropriately captures the difference of frames, it should be able to distinguish the difference of frames with a high match rate.

Table 3 lists the results of the average match rate. In the models of $\text{BERT}_{\text{LARGE}}$, RoBERTa, $\text{BERT}_{\text{BASE}}$, XLNet, and ALBERT, which obtained relatively high scores in the results shown in Table 2, we can see that the group without relations got higher scores than the group with relations. It is arguably concluded this result indicates that these models accurately captured the essential difference of frames.

5 Estimation of Number of Frames

In the experiments in Section 4, we gave the number of frames in our dataset to the Gaussian mixture model. However, it is necessary to estimate the number of frames that each verb evokes for semantic frame induction. Therefore, we investigated how well we can estimate the number of frames on the basis of information criteria by using contextualized word representations. Specifically, we adopted a Bayesian information criterion (BIC) (Schwarz, 1978), which is used for determining the number of clusters, and an adjusted-BIC, in which the BIC is adjusted so that the estimated number of clusters is close to the number of human-annotated frames.

5.1 Information Criterion

The Bayesian information criterion (BIC) is one of the most widely used criterion for model selection. The BIC is defined as

$$\text{BIC} = -2\ln(L) + k \cdot \ln(n_s), \qquad (1)$$

where L is the likelihood of the model, n_s is the number of samples, and k is the number of model parameters. The parameters of the Gaussian mixture model consist of the mean, covariance, and mixture weights. When the numbers of clusters and dimensions are represented by n_c and d, respectively, the number of parameters required to represent the mean is $d \times n_c$. Since we adopted spherical as the covariance type, where each component has its own single variance, the number of parameters required to represent the covariance is n_c. Since the mixture weights for each component are probabilities that sum to 1, the number of parameters required to represent the mixture weight is $n_c - 1$. Thus, the total number of model parameters is $k = (d + 2) \times n_c - 1$. When the BIC is used to determine the number of clusters, the number that minimizes the BIC is selected.

The first term on the right-hand side in Equation 1 decreases as the number of clusters increases because the likelihood of an optimized model generally increases as the number of parameters increases. The second term on the right-hand side is regarded as a penalty term that inhibits the increase in the number of clusters. The granularity of frames decided by human intuition may not be optimal in terms of the information criterion. Therefore, we introduce an adjusted-BIC in which the penalty term of the BIC is adjusted so that the granularity of the frames is close to human intuition. The equation of the adjusted-BIC (a-BIC) is defined as

$$\text{a-BIC} = -2\ln(L) + c \cdot k \cdot \ln(n_s), \qquad (2)$$

where c is a constant that adjusts the penalty, which is decided by using the development set.[11]

[10] Among the 120 verbs that were used as the test set in the experiment, the number of verbs that evoked only two frames was 96. The number of verbs in the group without the relations was 62, while the number of verbs in the group with the relations was 34.

[11] The constant c is increased from 1 in increments of 0.1, and we decide the value when the total number of frames and the total estimated number of clusters are as close as possible in the development set.

		ρ	Accuracy	RMSE
FrameNet	BIC	0.165	0.058	3.131
	a-BIC	0.177	0.517	1.195
PropBank	BIC	0.066	0.004	7.429
	a-BIC	0.631	0.608	1.008

Table 4: Evaluation on estimating the number of frames using BIC and a-BIC for FrameNet and PropBank.

	1	2	3	4	5+
1	1	16	32	19	52
2	0	11	14	19	52
3	0	1	1	5	13
4	0	0	0	1	3

BIC for FrameNet

	1	2	3	4	5+
1	0	0	1	10	109
2	0	1	1	2	81
3	0	0	0	1	11
4+	0	0	0	0	23

BIC for PropBank

	1	2	3	4	5+
1	88	17	7	4	4
2	50	35	5	3	3
3	10	10	0	0	0
4	2	1	0	1	0

a-BIC for FrameNet

	1	2	3	4	5+
1	80	35	3	0	2
2	15	57	12	0	1
3	1	2	7	1	1
4+	0	6	7	3	7

a-BIC for PropBank

Table 5: Confusion matrices using BIC and a-BIC for FrameNet and PropBank. Vertical axis represents the number of frames in our dataset, and horizontal axis represents the estimated number of clusters.

5.2 Results

In the experiments in Section 4, we used only the verbs that evoke more than one type of frame. However, it is also essential for verbs that evoke only one type of frame to recognize that. Therefore, we added verbs that evoke only one frame. The number of verbs added was the same as the number of verbs used in the experiment in Section 4. We also used a maximum of 100 annotated sentences from each frame. As a result, we used 116 verbs for parameter tuning as the development set and 240 verbs for evaluation as the test set for FrameNet, and we used 88 verbs for parameter tuning as the development set and 240 verbs for evaluation as the test set for PropBank.

We used $\text{BERT}_{\text{LARGE}}$ as contextualization word representations. We evaluated the automatic estimation of the number of frames by using Spearman's rank correlation coefficient (ρ), accuracy, and root mean square error (RMSE) for the estimated number of clusters and the number of frames in our dataset. Table 4 lists the estimation results of the number of frames. For both FrameNet and PropBank, using the adjusted-BIC as the information criterion resulted in better scores than using the BIC. When using the adjusted-BIC, Spearman's rank correlation coefficients were 0.177 and 0.631 for FrameNet and PropBank, respectively. The accuracy scores were over 0.5, which means that we could correctly predict the number of frames for more than half of the verbs. The accuracy for FrameNet is lower than the accuracy for PropBank. Accurate prediction of the number of frames for FrameNet will need to consider semantic coherence across different verbs, since frames in FrameNet are not defined independently for each verb.

Table 5 shows the confusion matrices between the number of human-annotated frames and the estimated number of frames using the BIC and the adjusted-BIC for FrameNet and PropBank. We can see that the BIC tended to overestimate the number of frames. The constant c was tuned at 3.1 for FrameNet and 3.4 for PropBank.

6 Conclusion and Future Work

We investigated to what extent contextualized word representations can recognize the difference of frames that the same verb evokes. Specifically, we focused on verbs that evoke multiple frames and performed clustering based on contextualized word representations of target verbs. We calculated the match rate between the generated clusters and the human-annotated frames and compared seven contextualized word representations: ELMo, $\text{BERT}_{\text{BASE}}$, $\text{BERT}_{\text{LARGE}}$, RoBERTa, ALBERT, GPT-2, and XLNet. We found that BERT, RoBERTa, XLNet, and ALBERT achieved high performance in distinguishing the difference of frames that the same verb evokes. We also found that we can estimate the number of frames with an accuracy of more than 50% by using the adjusted-BIC, which adjusts the penalty term of the BIC.

In this paper, we focused on the difference of frames that each verb evokes. That is, we analyzed each verb separately. However, in FrameNet, frames are shared by several verbs. For example, the verbs "support," "prove," and "demonstrate" can evoke the same EVIDENCE frame. To induce FrameNet-style frames, we need to investigate to what extent contextualized word representations capture frames over verbs.

Semantic frame induction requires not only distinguishing the difference of frames that the same verb evokes but also grouping its arguments by the semantic role. For example, if a sentence contains a verb that evokes the EVIDENCE frame, the sentence contains what is claimed and what supports the claim as its argument. Contextualized word representations of the arguments will also be useful

for grouping arguments by the same roles.

Furthermore, we only considered verbs as frame-evoking words, but we need to examine whether we can obtain similar results for words with other parts of speech that evoke frames such as nouns. These investigations are expected to bring us closer to our goal of automatically constructing high-quality semantic-frame resources. They can also induce semantic frames for under-resourced languages or specific domains since contextualized word representations do not require human-annotated texts.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers 18H03286 and 21K12012.

References

David J Aldous. 1985. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII—1983*, pages 1–198.

Asaf Amrami and Yoav Goldberg. 2018. Word sense induction with neural biLM and symmetric patterns. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, pages 4860–4867.

Saba Anwar, Dmitry Ustalov, Nikolay Arefyev, Simone Paolo Ponzetto, Chris Biemann, and Alexander Panchenko. 2019. HHMM at SemEval-2019 task 2: Unsupervised frame induction using contextualized word embeddings. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval 2019)*, pages 125–129.

Nikolay Arefyev, Boris Sheludko, Adis Davletov, Dmitry Kharchev, Alex Nevidomsky, and Alexander Panchenko. 2019. 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 2019)*, pages 31–38.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. 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 1998)*, pages 86–90.

Chris Biemann. 2006. 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*, pages 73–80.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Jackie Chi Kit Cheung and Gerald Penn. 2013. Towards robust abstractive multi-document summarization: A caseframe analysis of centrality and domain. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pages 1233–1242.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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 2019)*, pages 4171–4186.

Charles J Fillmore. 2006. Frame semantics. *Cognitive Linguistics: Basic Readings*, 34:373–400.

Aldo Gangemi, Diego Reforgiato Recupero, Misael Mongiovì, Andrea Giovanni Nuzzolese, and Valentina Presutti. 2016. Identifying motifs for evaluating open knowledge extraction on the web. *Knowledge-Based Systems*, 108:33–41.

Christian Hadiwinoto, Hwee Tou Ng, and Wee Chung Gan. 2019. Improved word sense disambiguation using pre-trained contextualized word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)*, pages 5297–5306.

Daisuke Kawahara, Daniel Peterson, Octavian Popescu, and Martha Palmer. 2014. Inducing example-based semantic frames from a massive amount of verb uses. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*, pages 58–67.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *Proceedings of the International Conference on Learning Representations (ICLR 2020)*.

Yang Liu. 2019. Fine-tune BERT for extractive summarization. *arXiv preprint arXiv:1903.10318*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Jiří Materna. 2012. Lda-frames: An unsupervised approach to generating semantic frames. In *International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2012)*, pages 376–387.

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

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. 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 2018)*, pages 2227–2237.

Behrang QasemiZadeh, Miriam R. L. Petruck, Regina Stodden, Laura Kallmeyer, and Marie Candito. 2019. SemEval-2019 task 2: Unsupervised lexical frame induction. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval 2019)*, pages 16–30.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Eugénio Ribeiro, Vânia Mendonça, Ricardo Ribeiro, David Martins de Matos, Alberto Sardinha, Ana Lúcia Santos, and Luísa Coheur. 2019. 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 2019)*, pages 130–136.

Gideon Schwarz. 1978. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Dan Shen and Mirella Lapata. 2007. Using semantic roles to improve question answering. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL 2007)*, pages 12–21.

Dmitry Ustalov, Alexander Panchenko, Andrey Kutuzov, Chris Biemann, and Simone Paolo Ponzetto. 2018. Unsupervised semantic frame induction using triclustering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pages 55–62.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS 2017)*, pages 5998–6008.

Bishan Yang and Tom Mitchell. 2017. A joint sequential and relational model for frame-semantic parsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 1247–1256.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems (NeurIPS 2019)*, pages 5754–5764.

Zheng Xin Yong and Tiago Timponi Torrent. 2020. Semi-supervised deep embedded clustering with anomaly detection for semantic frame induction. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 3509–3519.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tieyan Liu. 2019. Incorporating BERT into neural machine translation. In *Proceedings of the International Conference on Learning Representations (ICLR 2019)*.