Semantic Frame Induction using Masked Word Embeddings and Two-Step Clustering

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

Recent studies on semantic frame induction show that relatively high performance has been achieved by using clustering-based methods with contextualized word embeddings. However, there are two potential drawbacks to these methods: one is that they focus too much on the superficial information of the frame-evoking verb and the other is that they tend to divide the instances of the same verb into too many different frame clusters. To overcome these drawbacks, we propose a semantic frame induction method using masked word embeddings and two-step clustering. Through experiments on the English FrameNet data, we demonstrate that using the masked word embeddings is effective for avoiding too much reliance on the surface information of frame-evoking verbs and that two-step clustering can improve the number of resulting frame clusters for the instances of the same verb.

1 Introduction

Semantic frame induction is a task of mapping frame-evoking words, typically verbs, into semantic frames they evoke (and the collection of instances of words to be mapped into the same semantic frame forms a cluster). For example, in the case of example sentences from FrameNet (Baker et al., 1998) shown in (1) to (4) in Table 1, the goal is to group the examples into three clusters according to the frame that each verb evokes; namely, {(1)}, {(2)}, and {(3), (4)}. Unsupervised semantic frame induction methods help to automatically build high-coverage frame-semantic resources.

Recent studies have shown the usefulness of contextualized word embeddings such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) for semantic frame induction. For example, the top three methods (Arefyev et al., 2019; Anwar et al., 2019; Ribeiro et al., 2019) in Subtask-A of SemEval-2019 Task 2 (QasemiZadeh et al., 2019) perform clustering of contextualized word embeddings of frame-evoking verbs. However, these methods have two potential drawbacks.

(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)

Table 1: Example sentences of verbs "get" and "acquire" and frames that each verb evokes in FrameNet. (FRAME)

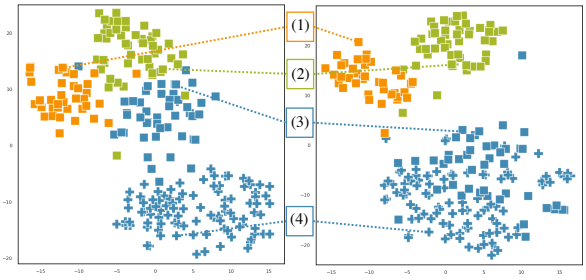

Figure 1: 2D projections of BERT embeddings of verbs (left) and masked verbs (right). Numbers in the figure correspond to numbers in Table 1, ■ and + are verbs "get" and "acquire", respectively, and each color indicates ARRIVING, TRANSITION_TO_STATE, and GETTING frame.

First, the contextualized word embeddings of the frame-evoking verbs strongly reflect the superficial information of the verbs. The left side of Figure 1 shows a 2D projection of contextualized embeddings of instances of the verbs "get" and "acquire" extracted from example sentences in FrameNet. Specifically, we extracted instances of "get" and "acquire" from FrameNet, obtained their embeddings by using a pre-trained BERT, and projected them into two dimensions by using t-distributed stochastic neighbor embedding (t-SNE) (Maaten and Hinton, 2008). As shown in the figure, among instances of "get", those that evoke the GETTING frame tend to be located close to instances of "acquire" that evokes the same GETTING frame. However, we can see that the difference be-

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Short Papers), pages 811–816
August 1–6, 2021.

tween verbs is larger than the difference between the frames that each verb evokes.

To remedy this drawback, we propose a method that uses a masked word embedding, a contextualized embedding of a masked word. The right side of Figure 1 shows a 2D projection of masked word embeddings for instances of the verbs "get" and "acquire". The use of masks can hide the superficial information of the verbs, and consequently we can confirm that instances of verbs that evoke the same frame are located close to each other.

The second drawback is that these methods perform clustering instances across all verbs simultaneously. Such clustering may divide instances of the same verb into too many different frame clusters. For example, if there are outlier vectors that are not typical for a particular verb, they tend to form individual clusters with instances of other frames in most cases. To solve this problem, we propose a two-step clustering, which first performs clustering instances of the same verb according to their meaning and then performs further clustering across all verbs.

2 Proposed Method

The proposed semantic frame induction method uses masked word embeddings and two-step clustering. We explain these details below.

2.1 Masked Word Embedding

A masked word embedding is a contextualized embedding of a word in a text where the word is replaced with a special token indicating that it has been masked, i.e., "[MASK]" in BERT. Our method leverages masked word embeddings of frame-evoking verbs in addition to standard contextualized word embeddings of frame-evoking verbs. In this paper, we consider the following three types of contextualized word embeddings.

v_{WORD}: Standard contextualized embedding of a frame-evoking verb.

v_{MASK}: Contextualized embedding of a frame-evoking verb that is masked.

$v_{\text{W+M}}$: The weighted average of the above two, which is defined as:

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

Here, $v_{\text{W+M}}$ is the weighted average of contextualized word embeddings with and without masking the frame-evoking verb. By properly setting the weight α using a development set, we expect to obtain embeddings that properly adjust the weight of superficial information of the target verb and information obtained from its context. $v_{\text{W+M}}$ is identical to v_{WORD} when α is set to 0 and identical to v_{MASK} when α is set to 1.

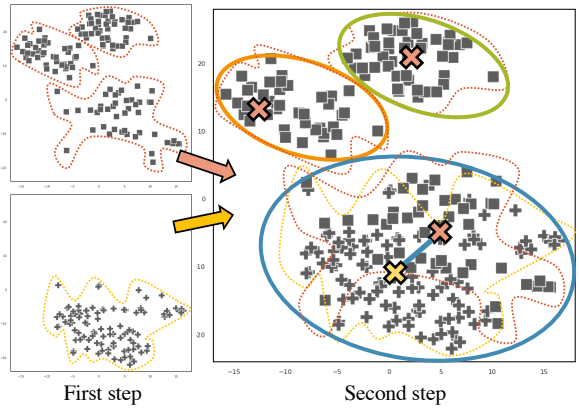

Figure 2: Flow of the two-step clustering. ■ and + denote the embeddings of "get" and "acquire", respectively.

2.2 Two-Step Clustering

In the two-step clustering, we first perform clustering instances of the same verb according to the semantic meaning and then perform further clustering across verbs. Finally, each generated cluster is regarded as an induced frame. Figure 2 shows the flow of the two-step clustering using the instances of "get" and "acquire" from FrameNet. As a result of the clustering in the first step, the instances of "get" are grouped into three clusters and the instances of "acquire" into one cluster. In the second step, one of the clusters of "get" and the cluster of "acquire" are merged. Consequently, three clusters are generated as the final clustering result. The details of each clustering are as follows.

Clustering Instances of the Same Verb The clustering in the first step aims to cluster instances of the same verb according to their semantic meaning. Since all the targets of the clustering are the same verbs, there should be no difference in the results between the cases using v_{WORD} and v_{MASK} as embeddings. Therefore, we use only v_{MASK} for this process. We adopt X-means (Pelleg and Moore, 2000) or group average clustering based on a Euclidean distance as the clustering algorithm.

While X-means automatically determine the number of clusters, group average clustering requires a clustering termination threshold. In the group average clustering, the distance between two

clusters is defined as the average distances of all instance pairs between clusters, and the cluster pairs with the smallest distance between clusters are merged in order. The clustering is terminated when there are no more cluster pairs for which the distance between two clusters is less than or equal to a threshold θ. In this study, θ is shared across verbs, not determined for each verb. Note that when θ is set to a sufficiently large value, the number of clusters is one for all verbs. To set θ to an appropriate value, we gradually decrease θ from a sufficiently large value and fix it to a value where the number of the generated frame clusters is equal to the actual number of frames in the development set.

In the theory of Frame Semantics (Fillmore, 2006) on which FrameNet is based, the association between a word and a semantic frame is called a lexical unit (LU). Since each cluster generated as the result of clustering in the first step is a set of instances of the same verb used in the same meaning, it can be considered to correspond to an LU. Therefore, we refer to it as pseudo-LU (pLU).

Clustering across Verbs The clustering in the second step aims to cluster the pLUs generated as the result of the first-step clustering across verbs according to their meaning. This step calculates average contextualized embeddings of each pLU and then clusters the pLUs by using the calculated embeddings across verbs. We adopt Ward clustering or group average clustering based on a Euclidean distance as the clustering algorithm.

We need a termination criterion for both clustering algorithms. A straightforward approach is to use the ratio of the number of frames to the number of verbs. However, this approach does not work well in this case since there is an upper limit to the number of frame types and the number of frames to be generated does not increase linearly with the number of verbs. Therefore, in this study, we use the ratio of pLU pairs belonging to the same cluster as the termination criterion. Specifically, the clustering is terminated when the ratio of pLU pairs belonging to the same cluster $p_{\mathrm{F}_1=\mathrm{F}_2}$ is greater than or equal to the ratio of LU pairs belonging to the same frame in the development set $p_{\mathrm{C}_1=\mathrm{C}_2}$. Here, $p_{\mathrm{F}_1=\mathrm{F}_2}$ is calculated as:

$$p_{\mathrm{F}_1=\mathrm{F}_2} = \frac{\text{\# of pLU pairs in the same cluster}}{\text{\# of all pLU pairs}}. \quad (2)$$

While the number of all pLU pairs is constant regardless of clustering process, the number of pLU pairs belonging to the same cluster monotonically increases as the clustering process progresses. $p_{\mathrm{C}_1=\mathrm{C}_2}$ can be calculated as well as $p_{\mathrm{F}_1=\mathrm{F}_2}$ and $p_{\mathrm{C}_1=\mathrm{C}_2}$ reaches 1 when the number of the entire cluster becomes one cluster. Therefore, $p_{\mathrm{C}_1=\mathrm{C}_2}$ is guaranteed to be greater than or equal to $p_{\mathrm{F}_1=\mathrm{F}_2}$ during the clustering process. Since the probability that randomly selected LU pairs belong to the same frame is not affected by the data size, the criterion is considered valid regardless of the data size.

	#Verbs	#LUs	#Frames	#Examples
Dev.	255	300	169	12,718
Test	1,017	1,188	393	47,499
All	1,272	1,488	434	60,217

Table 2: Statistics of the dataset from FrameNet.

3 Experiment

We conducted an experiment of semantic frame induction to confirm the efficacy of our method. In this experiment, the objective is to group the given frame-evoking verbs with their context according to the frames they evoke.

3.1 Setting

Dataset From Berkeley FrameNet data release 1.7[1] in English, we extracted verbal LUs with at least 20 example sentences and used their example sentences. That is, all target verbs in the dataset have at least 20 example sentences for each frame they evoke. We limited the maximum number of sentence examples for each LU to 100 and if there were more examples, we randomly selected 100. Note that we did not use the SemEval-2019 Task 2 dataset because the dataset is no longer available as described on the official web page.[2]

The extracted dataset contained 1,272 different verbs as frame-evoking words. We used the examples for 255 verbs (20%) as the development set and those for the remaining 1,017 verbs (80%) as the test set. Thus, there are no overlapping frame-evoking verbs or LUs between the development and test sets, but there is an overlap in the frames evoked. We divided the development and test sets so that the proportion of verbs that evoke more than one frames would be the same. The development set was used to determine the alpha of $v_{\mathrm{W+M}}$

[1] https://framenet.icsi.berkeley.edu/

[2] https://competitions.codalab.org/competitions/19159#learn_the_details-datasets

Model	Clustering		α	#pLU	#C	PU / iPU / PiF	BcP / BcR / BcF
1-cluster-per-head	1cpv		–	–	1017	88.9 / 39.7 / 54.9	86.6 / 33.9 / 48.7
Arefyev et al. (2019)	GA (Cosine)		–	–	995	69.9 / 55.1 / 61.6	62.8 / 44.0 / 51.7
Anwar et al. (2019)	GA (Manhattan)		–	–	891	71.5 / 52.0 / 60.2	65.1 / 41.0 / 50.3
Ribeiro et al. (2019)	Chinese Whispers		–	–	542	50.9 / 66.3 / 57.5	39.4 / 56.7 / 46.5
One-step clustering	Ward		0.0	–	393	64.3 / 49.5 / 56.0	55.2 / 38.9 / 45.6
	GA		0.0	–	393	38.7 / 64.9 / 48.5	26.1 / 52.5 / 34.9
Two-step clustering	**first-step**	**second-step**					
	1cpv'	Ward	0.8	1017	164	54.8 / 73.1 / 62.7	43.1 / 64.3 / 51.6
	1cpv'	GA	0.9	1017	412	69.0 / 71.3 / 70.1	60.5 / 62.3 / 61.4
	GA	Ward	0.9	1196	291	49.3 / 72.9 / 58.8	37.3 / 64.6 / 47.3
	GA	GA	0.6	1196	479	63.0 / **76.3** / 69.0	52.8 / **68.0** / 59.4
	X-means	Ward	0.8	1043	167	54.0 / 72.2 / 61.8	42.6 / 63.6 / 51.1
	X-means	GA	0.7	1043	410	**71.9** / 74.1 / **73.0**	**63.2** / 65.5 / **64.4**

Table 3: Experimental results. #pLU denotes the number of pLUs and #C denotes the number of frame clusters. Note that the actual numbers of LUs and frames are 1,188 and 393, respectively. GA means group average clustering.

and the termination criterion for the clustering in each step and layers to be used as contextualized word embeddings. Table 2 lists the statistics of the dataset.

Models We compared four models, all combinations of group average clustering or X-means in the first step and Ward clustering or group average clustering in the second step. We also compared a model that treats all instances of one verb as one cluster (1-cluster-per-verb; 1cpv) and models that treat all instances of one verb as one cluster (1cpv') in the first step and then perform the clustering in the second step.

In addition, we compared our models with the top three models in Subtask-A of SemEval-2019 Task 2. Arefyev et al. (2019) first perform group average clustering using BERT embeddings of frame-evoking verbs. Then, they perform clustering to split each cluster into two by using TF-IDF features with paraphrased words. Anwar et al. (2019) use the concatenation of the embedding of a frame-evoking verb and the average word embedding of all words in a sentence obtained by skip-gram (Mikolov et al., 2013). They perform group average clustering based on Manhattan distance by using the embedding. Ribeiro et al. (2019) perform graph clustering based on Chinese whispers (Biemann, 2006) by using ELMo embeddings of frame-evoking verbs.

To confirm the usefulness of the two-step clustering, we also compared our models with models that perform a one-step clustering. For the model, we used Ward clustering or group average clustering as the clustering method and $v_{\text{W+M}}$ as the contextualized word embedding. We gave the oracle number of clusters to these models, i.e., we stopped clustering when the number of human-annotated frames and the number of cluster matched.

Metrics and Embeddings We used six evaluation metrics: B-CUBED PRECISION (BCP), B-CUBED RECALL (BCR), and their harmonic mean, F-SCORE (BCF) (Bagga and Baldwin, 1998), and PURITY (PU), INVERSE PURITY (IPU), and their harmonic mean, F-SCORE (PIF) (Karypis et al., 2000). We used BERT (bert-base-uncased) in Hugging Face[3] as the contextualized word embedding.

3.2 Results

Table 3 shows the experimental results.[4] When focusing on BCF, which was used to rank the systems in Subtask-A of SemEval-2019 Task 2, our model using X-means as the first step and group average clustering as the second step achieved the highest score of 64.4. It also got the highest PIF score of 73.0. The number of human-annotated frames was 393, while the number of generated clusters was 410. These results demonstrate that the termination criterion of the two-step clustering works effectively.

In all two-step clustering methods, α was tuned between 0.0 and 1.0, which shows that both v_{WORD} and v_{MASK} should be considered. In addition, α was close to 1.0 for these methods, which indicates that v_{MASK} is more useful for clustering instances across verbs. In contrast, $v_{\text{W+M}}$ in the one-step clustering methods was equivalent to v_{WORD} with $\alpha = 0.0$. This indicates that there is no effect of using v_{MASK}

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

[4] The performance of the top three models in Subtask-A of SemEval-2019 Task 2 is lower than reported in the task because the dataset used in this study has a high proportion of verbs that evoke multiple frames and is, therefore, a challenging dataset.

for the one-step clustering-based methods.

The two-step clustering-based models that use group average clustering as the second clustering algorithm tended to achieve high scores. This indicates that the two-step clustering-based approach, which first cluster instances of the same verb and then cluster across verbs, is effective. However, as to the first clustering, 1cpv' strategy, which treats all the instances of the same verb as one cluster, achieved a higher accuracy than the clustering of the group average method, and achieved an accuracy close to the clustering of X-means, and thus we can say that 1cpv' strategy is effective enough for this dataset. We think this is due to the fact that the dataset used in this study is quite biased towards verbs that evoke only one frame, and we believe that the effectiveness of the 1cpv' may be limited in a more practical setting. Further investigation of this is one of our future works.

4 Conclusion

We proposed a method that uses masked word embeddings and two-step clustering for semantic frame induction. The results of experiments using FrameNet data showed that masked word embeddings and two-step clustering are quite effective for this frame induction task. We will conduct experiments in a setting where nouns and adjectives are also accounted for as frame-evoking words. The future goal of this research is to build a frame-semantic resource, which requires not only the induction of semantic frames but also the determination of the arguments required by each frame and the induction of semantic roles of the arguments. A possible extension of our approach is to utilize contextualized word embeddings of arguments of verbs to see if it is possible to generalize our approach for achieving this goal.

Acknowledgements

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

References

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.

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING 1998)*, pages 79–85.

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 (TextGraphs 2006)*, pages 73–80.

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.

Michael Karypis, Steinbach George, and Vipin Kumar. 2000. A comparison of document clustering techniques. In *Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining Workshop on Text Mining*.

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

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.

Dan Pelleg and Andrew Moore. 2000. X-means: Extending k-means with efficient estimation of the number of clusters. In *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pages 727–734.

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.

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.