Unsupervised Semantic Frame Induction Revisited

Younes Samih **Laura Kallmeyer**
Department of Computational Linguistics
Heinrich Heine University Düsseldorf,
Düsseldorf, Germany
{samih,kallmeyer}@hhu.de

Abstract

This paper addresses the task of semantic frame induction based on pre-trained language models (LMs). The current state of the art is to directly use contextualized embeddings from models such as BERT and to cluster them in a two step clustering process (first lemma-internal, then over all verb tokens in the data set). We propose not to use the LM's embeddings as such but rather to refine them via some transformer-based denoising autoencoder. The resulting embeddings allow to obtain competitive results while clustering them in a single pass. This shows clearly that the autoendocer allows to already concentrate on the information that is relevant for distinguishing event types.

1 Introduction

In natural language processing, Semantic Frame Induction refers to the task of clustering target word instances, specifically verbs, in a corpus according to their semantic frames in a given context. For example, in the sentences:

(a) *The price of LNG is rising, which makes the European economy unstable.*

(b) *Gold value fell 2% in Junuary after climbing 5% in August.*

(c) *Adam climbs dangerous cliffs.*

We would like to cluster the verbs in (a) and (b) in one group and (c) in another. The problem of verb semantic frames induction has received its share of attention, particularly in the SemEval 2019 shared task (Subtask-A) (QasemiZadeh et al., 2019a), in which the gold labels are annotated according to the FrameNet (Baker et al., 1998) frames inventory. Frame-semantic resources are prohibitively expensive and time-consuming to construct due to difficulties in the frame definitions, as well as the complexity of the construction and annotation tasks, that require expert knowledge in lexical event semantics. To overcome these issues, researchers proposed to automate the process of FrameNet construction through unsupervised techniques (Titov and Klementiev, 2011; Modi et al., 2012; Ustalov et al., 2018). Unsupervised semantic frame induction methods help to automatically build high-coverage frame-semantic resources. Up until recently, state-of-the-art results for semantic frame induction were dominated by a series of models leveraging contextualised pretrained language model representations to cluster instances of verbs according to the frames they evoke (Arefyev et al., 2019; Anwar et al., 2019). In recent work, Ribeiro et al. (2020) achieve state-of-the-art results by applying a graph-clustering algorithm based on Chinese whispers (Biemann, 2006) by using contextualized representations of frame-evoking verbs from BERT (Devlin et al., 2019)). Another approach has been proposed by Yamada et al. (2021b), who also use masked word embeddings and two-step clustering: each target instance is represented by three contextualized embeddings in a text, clustering is performed first over instances of the same verb and then across all verbs. However, these previous methods have one crucial shortcoming. As transformers based contextual words embeddings are originally designed to be fine-tuned on each downstream task to attain their optimal performance, it is unclear how best to extract representations of frame-evoking verbs from them, which are broadly applicable across diverse word-related tasks. In this paper, we further explore the use of LM representations by leveraging transformer-based Sequential Denoising Auto-Encoder (Wang et al., 2021) (TSDAE) embeddings to tackle the aforementioned problem. The proposed method achieves state-of-the-art performance on the frame induction task.

The contributions of this paper are three-fold:

- To the best of our knowledge, we are the first to adapt Transformer-based Sequential Denoising Auto-Encoder for semantic frame induction.
- Our method does not require two step-clustering, which is essential in most recent semantic frame induction models (Arefyev et al., 2019; Yamada et al., 2021a).
- Our clustering model outperforms recent state of the art systems for semantic frame induction on the SemEval 2019 shared task (Subtask-A) benchmark.

2 Method

In this section, we provide a brief description of TSDAE, and introduce the different components of our semantic frames induction model. The proposed model works in three stages:

- We train TSDAE on unlabeled sentences from the target task,
- then use its encoder to extract embeddings of the frame-evoking verb, associating each target verb instance with its representative vector.
- Finally, We perform clustering on these representative vectors.

TSDAE based word embeddings TSDAE, as shown in Figure 1, is a popular unsupervised learning algorithm based on an encoder decoder architecture. The model is a modified encoder-decoder Transformer where the key and value of the cross-attention are both restricted to yield sentence embedding only (Wang et al., 2021). The encoder maps the original input vector to a hidden representation, and the decoder maps the hidden representation back to the original input space. During training, noise is added to each input text by deleting or swapping a fraction of all tokens (we delete 60% of words in our experiments[1]), encoding the noisy text and reconstructing the embedding using the decoder module. The autoencoder minimizes the reconstruction error by approximating an identity function (Ng et al., 2011). A good reconstruction quality means that the semantics must be well captured in the word embeddings by the encoder. After training, the decoder module is discarded and the encoder is used to extract word representations. We re-implemented the TSDAE algorithm based on Huggingface's Transformers. [2] The algorithm is a simplified version of methods described in (Wang et al., 2021).

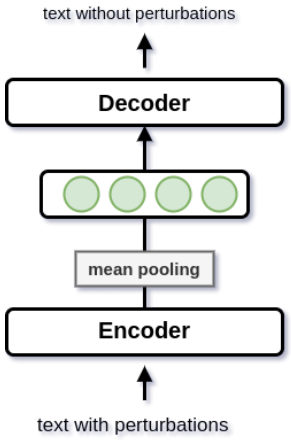

Figure 1: Architecture of TSDAE

Clustering After training TSDAE on unlabeled sentences from the target task, contextualised vectors for the frame evoking verbs in the sentences are calculated, and then clustered using agglomerative clustering with average linkage and cosine distance. The number of clusters is defined based on clusters maximizing the average silhouette score of all frame evoking verbs.

3 Experiments

Data	#Verbs	#Frames	#Examples
Dev.	600	41	588
Test.	4620	149	3346
All.	5220	190	3934

Table 1: Statistics of the dataset from the SemEval 2019 shared task

In this section, we introduce the datasets and experiment settings used for semantic frame induction. We also present the evaluation results of each model and compare them against existing semantic frame induction systems.

[1] We performed several auxiliary experiments on the development dataset to determine the optimal noise type and its ratio. It was discovered that removing the verb that evokes the semantic frame did not produce the most favorable outcome. Instead, setting the deletion ratio to 0.6 resulted in the most effective performance.

[2] https://github.com/huggingface/transformers

Model	Embeddings	#C	Pu	Ipu	Fpu	Bcp	Bcr	Bcf
1-cluster-per-verb	-	273	82.16	66.95	73.78	75.98	57.33	65.35
Anwar et al. (2019)	Elmo	150	72.4	81.49	76.68	62.17	75.27	68.1
Arefyev et al. (2019)	Bert	272	78.68	77.62	78.15	70.86	70.54	70.7
Ribeiro et al. (2019)	Bert	222	72.84	77.84	75.25	61.25	69.96	65.32
Ribeiro et al. (2020)	Bert	-	-	-	79.97	-	-	73.07
GA	Bert	227	80.26	79.05	79.65	73.52	71.88	72.69
	RoBERTa	192	80.35	81.9	81.12	73.61	75.7	**74.64**
TSDAE+GA	Bert	208	79.87	79.87	79.87	72.89	73.41	73.15
	RoBERTa	160	80.17	**84.33**	**82.2**	**73.62**	**78.67**	**76.06**

Table 2: Experimental results. #C denotes the number of frame clusters.Scores in bold denote significant improvements over the baseline. GA designates group average clustering.

3.1 Dataset

We use the SemEval 2019s Task 2 (QasemiZadeh et al., 2019b) as the benchmark datasets to evaluate our models and to facilitate comparison with related work. This dataset contains a subset of sentences extracted from the Penn Treebank 3.0 (Marcus et al., 1993) annotated with FrameNet Frames and tagged with morphosyntactic information in the CoNLL-U format (Buchholz and Marsi, 2006). Table 1 lists the statistics of the dataset.

3.2 Evaluation Measures

We evalute our approach using the six evaluation metrics[3] employed on the SemEval's task: Purity (Pu), inverse-Purity (Ipu) and their harmonic mean (Fpu) as proposed in (Steinbach et al., 2000), as well as The harmonic mean of BCubed's precision and recall (denoted by Bcp, Bcr, and Bcf respectively) (Bagga and Baldwin, 1998).

3.3 Experimental Settings

For all experiment we use BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019). In all cases we use the implementations from the HuggingFace Transformers toolkit (Wolf et al., 2019).

Baselines A very competitive baseline for frame-semantic induction is the SemEval'19 shared task 2 winning system by Arefyev et al. (2019). They use a two-step agglomerative clustering model. First, it groups examples to a relatively small number of large clusters, exploiting dense vector representations of the target word in a context obtained from hidden layers of BERT model. It merges verbs that evoke the same frame together while not taking into account homonyms. Then splits each of them into smaller clusters using the TF-IDF representations from substitutes generated for the target word by BERT masked LM to disambiguate all homonyms. An even stronger baseline is the system by Ribeiro et al. (2020) who apply Chinese whisper (Biemann, 2006), a graph-clustering algorithm, to a graph using contextualised representations of frame-evoking verbs as its nodes. Anwar et al. (2019) introduced a simpler system based on the agglomerative clustering of contextualised representations extracted from hidden layers of ELMo (Peters et al., 2018). Finally, we also evaluated one additional, simpler baseline (1-cluster-per-head) that treats all instances of one verb as one cluster.

3.4 Results

We evaluate the performance of each model and report the BCubed F1-scores in Table 2, along with the results from other semantic frame induction systems. Our model (TSDAE+GA) based on TSDAE and group average clustering outperforms the other methods on both Fpu and Bcf by a large margin. It achieves the highest Fpu score of 82.2 and also got the highest Bcf score of 76.06. The graph-based clustering by Ribeiro et al. (2020) proved to be the most competitive baseline, yielding decent scores according to all six measures. Finally, our RoBERTa group average model (GA) relying on hard clustering algorithms showed a slight increase in performance when compared to that of the graph-based model, justifying the more elaborate (TSDAE+GA) method. It is also worth noting that the Bert based group average model obtain a slightly worse or identical results.

[3]We use the standard evaluation script from the SemEval'19 shared task to calculate all the results. http://pars.ie/lr/semeval2019-task2/semeval-2019-task2-scorer.zip

4 Analysis

We extracted the cluster signatures and manually inspected all of the semantic frame clusters produced by TSDAE+GA, our best system, along with their associated verbs in order to scrutinize the emerging semantic classes and gain insight into annotator decisions. We found that the most prominent reason for incorrect clustering was due to the hard partitioning output, while the evaluation dataset contained fuzzy clusters. We also observed that the semantic distinctions that are easier for humans to make often elude representation models, and that discriminating between similar and highly associated but dissimilar verbs remains a challenge for most systems. Moreover, we noticed that the effectiveness of the models differ depending on the semantic Frames, indicating discrepancies in the quality of representations for verbs from diverse domains. Interestingly, we found that many clusters included an incoherent mix of multiple semantic frames along with an incoherent set of verbs. This suggests that frame induction should not be treated solely as a verb clustering task as it requires a distinct and separate approach.

5 Conclusion

In this paper, we introduced the first implementation of TSDAE for unsupervised frame induction and demonstrated that our method outperforms previous approaches in SemEval'19 shared task 2, setting a new state-of-the-art. Our error analysis revealed that a major source of incorrect clustering stemmed from the hard partitioning output, while the evaluation dataset consisted of fuzzy clusters. For future work, we aim to extend our work to the multi-lingual setup in future studies.

Acknowledgments

The work presented in this paper was partly financed by the Deutsche Forschungsgemeinschaft (DFG) within the project "Unsupervised Frame Induction (FInd)". We wish to thank three anonymous reviewers for their constructive feedback and helpful comments.

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*, pages 125–129, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

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*, pages 31–38, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

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, New York City. Association for Computational Linguistics.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 149–164, New York City. Association for Computational Linguistics.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Ashutosh Modi, Ivan Titov, and Alexandre Klementiev. 2012. Unsupervised induction of frame-semantic representations. In *Proceedings of the NAACL-HLT Workshop on the Induction of Linguistic Structure*, pages 1–7, Montréal, Canada. Association for Computational Linguistics.

Andrew Ng et al. 2011. Sparse autoencoder. *CS294A Lecture notes*, 72(2011):1–19.

Matthew E. 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, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

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

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

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*, pages 130–136, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Eugénio Ribeiro, Andreia Sofia Teixeira, Ricardo Ribeiro, and David Martins de Matos. 2020. Semantic frame induction through the detection of communities of verbs and their arguments. *Applied Network Science*, 5(1):1–32.

Michael Steinbach, George Karypis, and Vipin Kumar. 2000. A comparison of document clustering techniques.

Ivan Titov and Alexandre Klementiev. 2011. A Bayesian model for unsupervised semantic parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1445–1455, Portland, Oregon, USA. Association for Computational Linguistics.

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 (Volume 2: Short Papers)*, pages 55–62, Melbourne, Australia. Association for Computational Linguistics.

Kexin Wang, Nils Reimers, and Iryna Gurevych. 2021. TSDAE: Using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 671–688, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Kosuke Yamada, Ryohei Sasano, and Koichi Takeda. 2021a. Semantic frame induction using masked word embeddings and two-step clustering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 811–816, Online. Association for Computational Linguistics.

Kosuke Yamada, Ryohei Sasano, and Koichi Takeda. 2021b. Verb sense clustering using contextualized word representations for semantic frame induction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4353–4362, Online. Association for Computational Linguistics.