L^{2}F/INESC-ID at SemEval-2019 Task 2: Unsupervised Lexical Semantic Frame Induction using Contextualized Word Representations

Eugénio Ribeiro[1,2], **Vânia Mendonça**[1,2], **Ricardo Ribeiro**[1,3],
David Martins de Matos[1,2], **Alberto Sardinha**[1,2], **Ana Lúcia Santos**[4,5], **Luísa Coheur**[1,2]
[1] INESC-ID Lisboa, Portugal
[2] Instituto Superior Técnico, Universidade de Lisboa, Portugal
[3] Instituto Universitário de Lisboa (ISCTE-IUL), Lisboa, Portugal
[4] Centro de Linguística da Universidade de Lisboa, Portugal
[5] Faculdade de Letras da Universidade de Lisboa, Portugal
eugenio.ribeiro@l2f.inesc-id.pt, vania.mendonca@tecnico.ulisboa.pt

Abstract

Building large datasets annotated with semantic information, such as FrameNet, is an expensive process. Consequently, such resources are unavailable for many languages and specific domains. This problem can be alleviated by using unsupervised approaches to induce the frames evoked by a collection of documents. That is the objective of the second task of SemEval 2019, which comprises three subtasks: clustering of verbs that evoke the same frame and clustering of arguments into both frame-specific slots and semantic roles.

We approach all the subtasks by applying a graph clustering algorithm on contextualized embedding representations of the verbs and arguments. Using such representations is appropriate in the context of this task, since they provide cues for word-sense disambiguation. Thus, they can be used to identify different frames evoked by the same words. Using this approach we were able to outperform all of the baselines reported for the task on the test set in terms of Purity F_1, as well as in terms of BCubed F_1 in most cases.

1 Introduction

The Frame Semantics theory of language (Fillmore, 1976) states that one cannot understand the meaning of a word without knowing the context surrounding it. That is, a word may evoke different semantic frames depending on its context. Considering this relation, sets of frame definitions and annotated datasets that map text into the semantic frames it evokes are important resources for multiple Natural Language Processing (NLP) tasks (Shen and Lapata, 2007; Aharon et al., 2010; Das et al., 2014). The most prominent of such resources is the FrameNet (Baker et al., 1998), which provides a set of more than 1,200 generic semantic frames, as well as over 200,000 annotated sentences in English. However, this kind of resource is expensive and time-consuming to build, since both the definition of the frames and the annotation of sentences require expertise in the underlying knowledge. Furthermore, it is difficult to decide both the granularity and the domains to consider while defining the frames. Consequently, such resources only exist for a reduced amount of languages (Boas, 2009) and even English lacks domain-specific resources in multiple domains.

The problem of building semantic frame resources can be alleviated by using unsupervised approaches to induce the frames evoked by a collection of documents. The second task of SemEval 2019 aims at comparing unsupervised frame induction systems for building semantic frame resources for verbs and their arguments (Qasemi Zadeh et al., 2019). It is split into three subtasks. The first, Task A, focuses on clustering instances of verbs according to the semantic frame they evoke while the others focus on clustering the arguments of those verbs, both according to the frame-specific slots they fill, on Task B.1, and their semantic role, on Task B.2.

In this paper, we address the three subtasks by following an approach that takes advantage of the recent developments on the generation of contextualized word representations (Peters et al., 2018; Radford et al., 2018; Devlin et al., 2018). Such representations are able to disambiguate different word senses by varying the position of a word in the embedding space according to its context. This ability is important in the context of semantic frame induction, since different word-senses typically evoke different frames. To identify words that evoke the same frame or have the same role, our approach consists of clustering their representations by applying the Chinese Whispers algorithm (Biemann, 2006) to a similarity-based graph. This way, we do not need to define the number of clusters and there is no bias towards the

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 130–136
Minneapolis, Minnesota, USA, June 6–7, 2019.

generation of clusters of similar size.

In the remainder of the paper, we start by providing an overview of previous studies related to the task, in Section 2. Then, in Section 3, we describe our approach and explain how it differs from previous approaches. Section 4 describes our experimental setup. The results of our experiments are presented and discussed in Section 5. Finally, Section 6 summarizes the conclusions of our work and provides pointers for future work.

2 Related Work

Following the motivation described in the previous section, previous studies have employed unsupervised approaches for the induction of semantic frames and roles. However, most studies have focused on semantic role induction. For instance, Titov and Klementiev (2012) proposed two models based on the Chinese Restaurant Process (Ferguson, 1973). The factored model induces semantic roles for each predicate independently using an iterative clustering approach, starting with one cluster per argument. On the other hand, the coupled model takes into consideration a distance-dependent prior shared among different predicates. Arguments from different predicates are then used as vertices of a similarity graph and each argument selects another argument as a member of the same cluster based on that similarity. Overall, the coupled model performs slightly better than the factored one. In both cases, each argument is represented by a set of syntactic features – sentence voice, argument position, syntactic relation, and existing prepositions.

Lang and Lapata (2014) proposed a graph partitioning approach over a multilayer graph. Each layer corresponds to a feature, i.e., each pair of vertices (arguments) is connected through multiple edges, each corresponding to their similarity according to that feature. Then, two clustering approaches were considered, achieving similar results. The first is an adaptation of agglomerative clustering to the multilayer setting. Instead of combining the similarity values into a single score, it clusters the arguments in each layer and then combines the obtained scores into a multilayer score. Clusters with greater multilayer similarity are then merged together, with larger clusters being prioritized. The second clustering approach consists of propagating cluster membership along the graph edges. In both cases, the combination of the scores of each layer is based on a set of conditions, in order to avoid having to learn or guess weights for each feature.

In contrast to the previous approaches, Titov and Khoddam (2015) proposed a reconstruction-error maximization framework which comprises two main components: an auto-encoder, responsible for labeling arguments with induced roles, and a reconstruction model, which takes the induced roles and predicts the argument that fills each role, i.e., it tries to reconstruct the input. The learning error is obtained by comparing the reconstructed argument to the original one. This enables the use of a larger feature set and more complex features, similarly to supervised approaches.

Concerning frame induction, Ustalov et al. (2018a) proposed a graph-based approach for the triclustering of Subject-Verb-Object (SVO) triples extracted using a dependency parser. Each vertex in the graph is the SVO triple, represented by the concatenation of word embeddings for the three elements. Vertices are connected to their k-nearest neighbours (k=10) according to their cosine similarity. The clusters are then generated using the Watset fuzzy graph clustering algorithm (Ustalov et al., 2017), which induces word-sense information in the graph before clustering. For each cluster, the corresponding triframe is generated by aggregating the subjects, verbs, and objects into separate sets and generating a triple using those sets. This approach outperformed hard clustering approaches, as well as topic-based approaches, such as LDA-Frames (Materna, 2012).

3 Induction Approach

Considering the subtasks we are approaching, we must use an approach that is able to induce not only semantic roles, but also semantic frames and its slots. In this sense, of the approaches described in the previous section, the triclustering approach proposed by Ustalov et al. (2018a) is the only one able to induce frames. However, in the context of our task, it has two major flaws. First, it focuses on the clustering of SVO triples, i.e., a frame is defined by a head and two slots. In our case, each instance has a variable number of arguments. Thus, the triclustering approach is not appropriate. Furthermore, since the arguments are clustered in combination with the verb, this approach is particularly inappropriate for semantic role induction. The second flaw is related to the approach

used for inducing word-sense information, which requires a thesaurus to provide synonymity information. Such resources must be manually built and, thus, may not be available for every language or lack domain-specific information.

We approach the first flaw by clustering the verb and its arguments independently. This way, we are able to cluster the instances of verbs to identify the frame heads, as required for Task A, and the instances of arguments to identify semantic roles, as required for Task B.2. To identify the slots of each frame, as required for Task B.1, we combine the clusters of the verbs with those of the arguments.

To deal with the second flaw, we replace the per-word embeddings used by Ustalov et al. (2018a) with contextualized word representations. These include information concerning the context in which a word appears and, thus, the position of a word in the embedding space varies according to that context. By using such representations, we are able to discard the fuzzy clustering approach used by Ustalov et al. (2018a) to induce word-sense, since it is revealed by the contextual variations of the representation of a word. Therefore, a hard clustering algorithm can be applied directly.

Algorithm 1 Induction Approach

Input: T // The set of head tokens to cluster
Input: EMBED // The contextualized embedding approach
Input: THRESH // The function for computing the neighboring threshold
Output: C // The set of clusters

1: $V \leftarrow \{\textsc{Embed}(t) : t \in T\}$ // The whole sentence is required for embedding generation
2: $D \leftarrow \{1 - \cos(\theta_{v,v'}) : (v, v') \in V^2, v \neq v'\}$ // $\theta_{v,v'}$ is the angle between the two vectors
3: $t \leftarrow \textsc{Thresh}(D)$
4: $E \leftarrow \{(v, v', D_{v,v'}) : (v, v') \in V^2, v \neq v', D_{v,v'} < t\}$ // The edge is weighted with the cosine distance between the vertices
5: $C \leftarrow \textsc{ChineseWhispers}(V, E)$
6: **return** C

Our approach is summarized in Algorithm 1. It starts by generating the contextualized representation of each instance to be clustered. In cases where the verb or argument to cluster consists of multiple words, we use a dependency parser to identify the head word and use its contextualized representation, since it contains information from the other words. Then, in order to build a graph, we compute the pairwise distances between the instances. These distances are used to decide which instances are considered neighbors. Since each instance is represented as a vector in the embedding space, we use the cosine distance. Moreover, since using a fixed number of neighbors is not realistic, we decided to use a threshold based on this distance. This threshold defines the granularity of the clusters and varies according to the set of instances. Instead of using a fixed threshold, we define it based on the parameters of the pairwise distances distribution. The actual combination of the parameters varies according to the subtask and is further discussed in the subsections below. Finally, to obtain the clusters, we apply the Chinese Whispers algorithm (Biemann, 2006) on a graph where the vertices are the instances and the edges connect neighbor instances. The weight of each edge is given by the distance between neighbors. We use the Chinese Whispers algorithm since it chooses the number of clusters on its own and is able to handle clusters of different sizes, thus being appropriate for the task. Furthermore, it has been proved successful in NLP clustering tasks.

3.1 Verb Clustering

The first subtask focuses on clustering verbs that evoke the same frame. The number of frames evoked in a set of documents is typically larger than the number of semantic roles and even larger in comparison to the number of slots per frame. Thus, a lower neighboring threshold is required to achieve such granularity. In our experiments, we achieved the best results when defining the neighboring threshold for clustering verbs, t_f, as

$$t_f = \frac{\mu + \sigma}{2}, \tag{1}$$

where μ and σ are the mean and standard deviation of the pairwise distance distribution, respectively. Using this threshold may lead to the induction of frames with different granularity, depending on the sense similarity between the verbs present in the dataset. However, if the induced frames are considered too abstract, the approach can be applied hierarchically on the instances of each cluster to obtain finer-grained frames.

3.2 Argument Clustering

Both the second and third subtasks focus on clustering arguments. However, while the second focuses on doing so in a per-frame manner to induce

its slots (frame elements), the third focuses on clustering them independently of the frame, i.e., to induce generic semantic roles. In the first case it would make sense to cluster the arguments of verbs that evoke each frame independently of the others. However, that may not be feasible on small datasets. Thus, we opted for clustering all the arguments together in both cases. The slot clusters for the second subtask are then given by the combination of the verb and argument clusters. Thus, this approach considers that slots are per-frame specializations of the semantic roles, which is accurate in most situations.

As previously stated, the number of semantic roles is typically smaller than the number of frames. Thus, a higher neighboring threshold can be used. In our experiments, we achieved the best results when defining the neighboring threshold while clustering arguments, t_a, as

$$t_a = \mu - 1.5\sigma. \quad (2)$$

Finally, since the arguments are highly dependent on the verb, we also performed experiments in which we combined the contextualized representation of the argument with that of the verb before applying the clustering approach.

4 Experimental Setup

In this section we describe our experimental setup in terms of data, implementation details, and evaluation metrics and baselines.

4.1 Dataset

In our experiments, we used the dataset provided by the task organization, built with sentences from the Penn Treebank 3.0 (Marcus et al., 1993), and annotated with FrameNet frames (Task A), frame elements or slots (Task B.1) and generic semantic roles (Task B.2). The development set consists of 600 verb-argument instances, 588 sentences and 1,211 arguments. The (blind) test set comprises 4,620 verb-argument instances, 3,346 sentences, 9,466 arguments labeled for semantic role and 9,510 arguments labeled for frame slot. Additionally, morphosyntactic information is provided in the CoNLL-U format (Buchholz and Marsi, 2006).

4.2 Implementation Details[1]

In our experiments we compared the performance of two approaches to generate the contextualized word representations. The first, ELMo (Peters et al., 2018), is based on bi-directional LSTMs (Hochreiter and Schmidhuber, 1997) and was the first approach to generate contextualized representations. Its output provides a context-free representation of the word and context information at two levels. In our experiments we use the sum of all information, since it leads to variations of the context-free representation according to the context. The second representation, BERT (Devlin et al., 2018), is based on the Transformer architecture (Vaswani et al., 2017) and currently leads to state-of-the-art results on multiple benchmark NLP tasks. Its output can be extracted from a single layer or the multiple layers included in the model. Contrarily to the ELMo layers, these do not have an associated semantics. Thus, we use the output of the last layer, since it contains information from all that precede it. In both cases we used pre-trained models. To obtain embedding vectors with the same dimensionality, 1,024, we used the ELMo model provided by the AllenNLP package (Gardner et al., 2017) and the large uncased BERT model provided by its authors.

To apply the Chinese Whispers algorithm, we relied on the implementation by Ustalov et al. (2018b), which requires the graph to be built using the NetworkX package (Hagberg et al., 2004). We did not use weight regularization and performed a maximum of 20 iterations. Furthermore, in order to avoid result changes based on non-deterministic factors, we fixed the random seed as 1337.

Finally, to obtain the syntactic dependencies used to determine the head token of multi-word verbs or arguments, we used the annotations provided with the task dataset.

4.3 Baselines

For comparison purposes, in addition to our results, we report the baselines provided by the task scorer. For the frame induction subtask (Task A), the baseline consists of assigning each verb lemma to a frame (*Lemma*). For the semantic role induction subtask (Task B.2), arguments are assigned to clusters according to their syntactic relation to the head verb (*Dep*). For the frame slot induction subtask (Task B.1), the previous baselines are combined by assigning each pair of verb lemma and argument's syntactic dependency to a cluster (*Lemma + Dep*). On the test set, we also consider a random assignment to the gold number of clus-

[1] https://gitlab.l2f.inesc-id.pt/eugenio/find/

	Approach	#C	Purity	inv-Purity	Purity F_1	B^3 Precision	B^3 Recall	B^3 F_1
Task A	**ELMo**	32	93.17	96.50	**94.80**	89.06	95.63	**92.23**
	BERT	72	89.67	84.00	86.74	83.17	77.77	80.38
	BL: Lemma	35	93.50	85.67	89.41	90.22	79.63	84.60
Task B.1	**ELMo**	72	68.35	72.40	70.31	57.60	64.18	60.72
	BERT	170	52.98	72.98	61.39	46.82	62.77	53.64
	ELMo + Verb	72	68.35	72.40	70.31	57.60	64.18	60.72
	BL: Lemma + Dep	136	84.30	70.74	**76.93**	78.71	58.36	**67.03**
Task B.2	**ELMo**	11	62.23	69.01	65.44	46.75	56.33	51.10
	BERT	72	48.35	83.97	61.36	38.94	72.86	50.75
	ELMo + Verb	140	70.17	43.80	53.93	62.20	23.27	33.87
	Dep + PoS	66	65.95	29.26	40.53	55.61	20.05	29.47
	BL: Dep	22	67.93	71.32	**69.59**	53.31	57.67	**55.41**

Table 1: Results obtained on the development set. The baselines are identified with *BL*.

ters as a baseline. Due to space constraints, we do not report the results of the remaining baselines proposed by Kallmeyer et al. (2018).

We report the results of an additional baseline for Task B.2 which considers both the argument's syntactic relation to the head verb and its Part-of-Speech (POS) tag (*Dep + POS*).

4.4 Evaluation metrics

We report our results using the metrics defined for the task: number of clusters (*#C*), purity, inverse-purity, and their harmonic mean (Purity F_1), as proposed by Steinbach et al. (2000), and BCubed (B^3) precision, recall, and F_1, as proposed by Bagga and Baldwin (1998).

5 Results

The results obtained on the development set are reported in Table 1. We can see that using ELMo to obtain the contextualized word representations leads to better results than BERT on every subtask. This is somewhat surprising since BERT is the state-of-the-art approach to generate contextualized representations. A possible explanation may lie in the fact that the two levels of ELMo which provide context information can be related to syntax and semantics (Peters et al., 2018), making them highly related to the task. On the other hand, the information provided by BERT representations is not as easy to categorize. Moreover, in every case, the number of clusters is underestimated when using ELMo and overestimated when using BERT.

On the frame induction subtask (Task A), our approach surpasses every baseline, but only when using ELMo embeddings. The lemma baseline is surpassed by over 5 percentage points on Purity F_1 and 7.5 on BCubed F_1. The same is not true on the other tasks, with the clustering based on the dependency relation between the argument and verb achieving the best results. It outperforms our approach in terms of both F_1 metrics by around 6.5 percentage points on the slot induction subtask (Task B.1) and around 4 points on the semantic role induction subtask (Task B.2). We believe that this happens because the development set is small and the kind of arguments does not vary much.

Combining the verb representation with that of the argument leads to worse results on Task B.2, since it is clustering the semantic roles per verb. On Task B.1, the result is the same as without using the verb representation, which suggests that the information provided by the verb is not able to improve the induced slots, but only to attribute them to the corresponding frame.

The approach which combines the dependency relation with the POS tag obtains worse results on Task B.2, as it leads to additional partitioning of the clusters. Thus, a large number of clusters is generated, which is not consistent with the nature of semantic roles.

The results obtained on the test set are reported in Table 2. We only submitted the clusters obtained using ELMo, since it outperformed BERT on the development set. Similarly, we did not consider the combination of verb and argument representation for the argument clustering tasks. However, we assessed the performance of the baseline based on the dependency relation and the POS tag.

On Task A, our approach surpasses all the baselines in terms of Purity F_1, but by less than 2 percentage points. In fact, it has a similar perfor-

	Approach	#C	Purity	inv-Purity	Purity F_1	B^3 Precision	B^3 Recall	B^3 F_1
Task A	**ELMo**	222	72.84	77.84	**75.25**	61.25	69.96	65.32
	BL: Lemma	273	82.16	66.95	73.78	75.98	57.33	**65.35**
	BL: Random	**149**	15.30	5.74	8.34	6.82	3.85	4.92
Task B.1	**ELMo**	526	58.26	64.30	**61.13**	44.79	53.21	**48.64**
	BL: Lemma + Dep	1203	78.46	45.99	57.99	71.11	33.77	45.79
	BL: Random	**436**	11.25	6.09	7.90	6.07	4.82	5.37
Task B.2	**ELMo**	6	58.29	71.19	**64.10**	36.80	60.15	**45.66**
	Dep + PoS	159	57.39	26.25	36.03	41.41	15.07	22.1
	BL: Dep	37	61.44	51.53	56.05	40.89	37.33	39.03
	BL: Random	**32**	34.77	4.85	8.51	21.92	3.46	5.98

Table 2: Results obtained on the test set. The baselines are identified with *BL*.

mance to the lemma baseline in terms of BCubed F_1. This happens because it overestimates the number of clusters, which suggests that the problem may be related to the threshold. However, using a threshold that leads to the induction of a number of frames similar to the gold standard ends up generating clusters of lower quality. This suggests that additional features must be introduced.

On the remaining tasks, our approach performs better than every baseline, which supports the claim that the better performance of the clustering approach based on the dependency relation on the development set is due to the limited variation in the kinds of argument present in that set. We observed an improvement of around 4 percentage points on Task B.1 on both F_1 metrics, and above 8 percentage points on Purity F_1 and nearly 7 on BCubed F_1 on Task B.2.

Once again, the approach which combines the dependency relation with the POS tag leads to worse results on Task B.2, due to additional partitioning of the clusters. In this case, the number of semantic roles is even more overestimated.

6 Conclusions

In this paper we presented our approach on unsupervised semantic frame, slot, and role induction in the context of the second task of SemEval 2019. The approach is based on the clustering of contextualized word representations of verbs and arguments. Using such representations is appropriate for the task since they provide word-sense information which is important for distinguishing the evoked frames.

We were able to achieve results that surpassed or performed on par with every baseline proposed for the three subtasks on the test set. However, the results are far from perfect and below those achieved by more complex approaches on the task, which suggests that the contextualized representations on their own are not able to provide all the information required to perform an accurate frame induction. Thus, as future work, we intend to assess the cases that our approach fails to cluster, and introduce additional features that provide relevant information for those cases, either by using a weighted combination of per-feature distance functions or a multilayer graph similar to that proposed by Lang and Lapata (2014).

Furthermore, since the number of instances in the test set is larger than in the development set, it may be feasible to apply a per-frame clustering approach for the slot induction task. This way, the induced slots are no longer mere specifications of the generic semantic roles.

Finally, although the number of semantic roles is not consensual in the literature, there is a set of core semantic roles which is common to every theory. Thus, it would be interesting to take advantage of that information to apply clustering approaches with a pre-defined number of clusters for semantic role induction. In fact, it would be interesting to explore other clustering approaches on every task and compare their performance with that of the Chinese Whispers algorithm.

Acknowledgements

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2019. Vânia Mendonça is funded by an FCT grant with reference SFRH/BD/121443/2016. The use of the corpus was licensed by the Linguistic Data Consortium (LDC).

References

Roni Ben Aharon, Idan Szpektor, and Ido Dagan. 2010. Generating Entailment Rules from Framenet. In *ACL*, volume 2, pages 241–246.

Amit Bagga and Breck Baldwin. 1998. Algorithms for Scoring Coreference Chains. In *LREC*, pages 563–566.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *ACL/COLING*, volume 1, pages 86–90.

Chris Biemann. 2006. Chinese Whispers: An Efficient Graph Clustering Algorithm and its Application to Natural Language Processing Problems. In *Workshop on Graph-based Methods for Natural Language Processing*, pages 73–80.

Hans C. Boas, editor. 2009. *Multilingual FrameNets in Computational Lexicography: Methods and Applications*. Mouton de Gruyter.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X Shared Task on Multilingual Dependency Parsing. In *CoNLL*, pages 149–164.

Dipanjan Das, Desai Chen, André F. T. Martins, Nathan Schneider, and Noah A. Smith. 2014. Frame-Semantic Parsing. *Computational Linguistics*, 40(1):9–56.

Jacob Devlin, Ming-wei Chang, Lee Kenton, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.04805.

Thomas S. Ferguson. 1973. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):209–230.

Charles J. Fillmore. 1976. Frame Semantics and the Nature of Language. *Annals of the New York Academy of Sciences*, 280(Origins and Evolution of Language and Speech):20–32.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. AllenNLP: A Deep Semantic Natural Language Processing Platform. *CoRR*, abs/1803.07640.

Aric Hagberg, Dan Schult, and Pieter Swart. 2004. NetworkX. GitHub.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Laura Kallmeyer, Behrang Qasemi Zadeh, and Jackie Chi Kit Cheung. 2018. Coarse Lexical Frame Acquisition at the Syntax–Semantics Interface Using a Latent-Variable PCFG Model. In *SEM*, pages 130–141.

Joel Lang and Mirella Lapata. 2014. Similarity-Driven Semantic Role Induction via Graph Partitioning. *Computational Linguistics*, 40(3):633–670.

Mitchell Marcus, Beatrice Santorini, and Mary Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):330–331.

Jiří Materna. 2012. LDA-Frames: An Unsupervised Approach to Generating Semantic Frames. In *CICLing*, pages 376–387.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *NAACL-HLT*, volume 1, pages 2227–2237.

Behrang Qasemi Zadeh, Miriam R L Petruck, Stodden Regina, Laura Kallmeyer, and Marie Candito. 2019. Semeval 2019 task 2: Unsupervised lexical frame induction. In *SemEval@NAACL-HLT*. The Association for Computer Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. Preprint.

Dan Shen and Mirella Lapata. 2007. Using Semantic Roles to Improve Question Answering. In *EMNLP-CoNLL*, pages 12–21.

Michael Steinbach, George Karypis, and Vipin Kumar. 2000. A Comparison of Document Clustering Techniques. In *KDD Workshop on Text Mining*.

Ivan Titov and Ehsan Khoddam. 2015. Unsupervised Induction of Semantic Roles within a Reconstruction-Error Minimization Framework. In *NAACL-HLT*, volume 1, pages 1–10.

Ivan Titov and Alexandre Klementiev. 2012. A Bayesian Approach to Unsupervised Semantic Role Induction. In *EACL*, volume 1, pages 12–22.

Dmitry Ustalov, Alexander Panchenko, and Chris Biemann. 2017. Watset: Automatic Induction of Synsets from a Graph of Synonyms. In *ACL*, volume 1, pages 1579–1590.

Dmitry Ustalov, Alexander Panchenko, Andrei Kutuzov, Chris Biemann, and Simone Paolo Ponzetto. 2018a. Unsupervised Semantic Frame Induction using Triclustering. In *ACL*, volume 2, pages 55–62.

Dmitry Ustalov et al. 2018b. Chinese Whispers for Python. GitHub.

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 *NIPS*, pages 5998–6008.