Joint Identification and Induction of Semantic Frames with Scalable Semi-Supervised Graph Clustering

Fabian Barteld[1,*] **Steffen Remus**[2,*] **Saba Anwar**[2] **Julian Stawecki**[1]
Alexander Ziem[1] **Chris Biemann**[2]

[1]Department of German Linguistics
Institute of German Studies
Heinrich-Heine-Universität Düsseldorf
firstname.lastname@uni-duesseldorf.de

[2]Language Technology Group
Department of Informatics
Universität Hamburg
firstname.lastname@uni-hamburg.de

Abstract

Current methods for automatically assigning frames to their evoking words can be divided into frame identification and frame induction. In frame identification, frame names coming from a labeled dataset are assigned to unseen instances, a classical supervised labeling task. However, the training datasets are known to be incomplete in terms of real-world frames, resulting in an issue with potentially new frame labels. In frame induction, instances are clustered regarding the frames they evoke, a classical unsupervised clustering task. However, existing training data is not used to identify known frames. To overcome these shortcomings, we propose to use semi-supervised clustering for combined frame identification and frame induction. By using constrained clustering with hard constraints coming from labeled data, the resulting clusters contain only labeled instances with the same label. Thus, frame names can be easily assigned. We show for English and German datasets that using semi-supervised clustering improves the quality of frame induction compared to unsupervised clustering methods and results in notably good performance regarding frame identification.

Keywords: Frame identification, Frame induction, Semi-Supervised Clustering

1. Introduction

Frame semantics is a semantic theory, assuming that meaning-bearing words or phrases evoke semantic frames, i.e., broadly speaking, general knowledge about situations and participants (Fillmore, 1982). Pioneered by the Berkeley FrameNet (BFN) (Baker et al., 1998) for English, such frames and their evoking elements are collected in so-called FrameNets which have since been created for many languages.[1] However, such manually created FrameNets have a coverage problem (Palmer and Sporleder, 2010), i.e., the frames contained in a FrameNet, like the BFN, only cover a fraction of the frames evoked in texts. Regarding BFN, Tsujimoto et al. (2025) also argue for a representativeness problem, i.e., the frequencies of frame-evoking elements (FEEs), also called targets, and frames in a FrameNet differ from the frequencies found in real-world corpora. In this paper, we present an automatic approach for extending a FrameNet with data from large-scale real-world corpora, working towards a solution for both problems.

Automatic approaches for applying frames to FEEs in sentences of a corpus can be divided into two basic settings: *Frame Identification (FId)*, where the frames invoked by the FEEs are identified based on a given list of frames and *Frame Induction (FIn)*, where targets evoking the same frame are grouped together. While FId is commonly treated as a supervised labeling problem, FIn is commonly treated as an unsupervised clustering problem (cf. Section 2).

However, both approaches have limitations: FId, on one hand, is known to suffer from the coverage problem described above (Sikos and Padó, 2019). This makes the application of FId to large corpora tedious. Many targets will necessarily be mislabeled since the correct label is not part of the training data. FIn, on the other hand, treated as an unsupervised problem, ignores existing knowledge about frames from FrameNets.

To overcome these shortcomings, we propose to apply semi-supervised clustering (SSC) to combine (unsupervised) FIn and (supervised) FId. Targets evoking known frames are labeled by this approach, and at the same time, targets evoking unknown frames are clustered into groups evoking the same, unnamed frame. As such, our approach provides a helpful resource to enrich a FrameNet with new instances for existing frames and with new frames from real-world corpora.

For SSC, we adapt Chinese whispers (CW) (Biemann, 2006), a scalable and efficient graph clustering algorithm which has previously been successfully used for unsupervised FIn (Ribeiro et al., 2020) and use it to cluster contextualized embed-

[*]Equal contribution.

[1]See for example the list at https://framenet.icsi.berkeley.edu/framenets_in_other_languages (visited October 2025).

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 10020–10030
11–16 May 2026.

dings of targets. We apply this clustering algorithm in a local-global setup, where the instances for individual verbs are first clustered separately and are merged afterwards. This increases the scalability of the overall approach.

We evaluate this joint classification and induction method using the English BFN dataset (Baker et al., 1998) in version 1.7, and the German SALSA dataset (Burchardt et al., 2006) in version 2.0 (Rehbein et al., 2012) in standard settings for both FId and FIn.[2]

2. Related Work

In this section, we present approaches to FId and FIn as well as SSC from the literature.

Frame Identification is a multiclass classification task where, given a sentence and a specific target word, the objective is to identify the exact frame evoked by this target. Frame identification has been addressed both, as a standalone task (Sikos and Padó, 2019; Popov and Sikos, 2019; Tan and Na, 2019; Jiang and Riloff, 2021; Su et al., 2021; Tamburini, 2022), and as a sub-task in a more complex task of semantic frame parsing, that additionally performs target identification and semantic role labeling tasks either independently as a pipeline approach (Das et al., 2010; Hermann et al., 2014; Hartmann et al., 2017; Swayamdipta et al., 2017) or in a joint fashion (Yang and Mitchell, 2017). A wide range of approaches has been used in previous work, which includes simple feature engineering-based probabilistic models (Das et al., 2010), distributed word representations of targets and their context (Hermann et al., 2014; Hartmann et al., 2017), pre-trained target embeddings combined with learned embeddings of their POS tags and token types (Swayamdipta et al., 2017), and pre-trained transformer-based models (Sikos and Padó, 2019; Tan and Na, 2019).

More recent work has not limited itself to the labeled sentences contained in BFN but successfully added further information like the relationships between frames (Popov and Sikos, 2019) and the frame descriptions (Su et al., 2021; Tamburini, 2022) using graph embeddings.

While these discussed approaches differ vastly, they all treat FId as a labeling task with a fixed set of labels (cf. the definition given in Su et al., 2021) and cannot deal with unknown frames. One approach that handles unknown frames is presented by Yong and Torrent (2020).[3] In this approach, a lexical unit (LU), essentially a word sense in FrameNet terminology, is first tagged with a label whether it evokes a known frame or not. In a second step, LUs evoking a known frame are labeled with the respective frame. However, the approach does not deal with targets on the instance level but with LUs. Furthermore, LUs not evoking a known frame are not further clustered.

So, for all these approaches, what QasemiZadeh et al. (2019) stated in their motivation for the unsupervised lexical frame induction task about semantic parsing with FrameNet data remains relevant: "these methods cannot extend beyond previously seen training labels, tagging out-of-domain semantics as unknown at best. This limitation does not hinder unsupervised methods, which will port and extend the coverage of semantic parsers" (p. 17).

Frame Induction has the goal to automatically discover meaningful frames from a text corpus. The key idea is to capture the underlying relationships between words that frequently appear in the same contexts and organize them into groups. Unsupervised clustering approaches are therefore suitable for this purpose.

Early works on FIn, such as Green et al. (2004), used manually crafted knowledge resources like dictionaries and WordNet (Fellbaum, 1998) to represent frame-evoking verbs with TF-IDF-like feature vectors and applied agglomerative clustering to find groups of verbs evoking the same frame.

Ustalov et al. (2019), on the other hand, presented a knowledge-free approach that performs fuzzy clustering of words in two hard clustering stages by first resolving sense ambiguities of a word by using a symbolic neighborhood graph as input, i.e., words and their most related collocates within a large corpus of text, such as Wikipedia, and creating a so-called *intermediate sense graph* using vector comparison for further disambiguating interrelated words. Each subgraph of similarities that spans a particular word is then clustered into different senses, such that each neighbor of a word is assigned to a particular cluster; this is called the *local step*. For the so-called *global step*, the local cluster elements are disambiguated to form a global graph of senses instead of words. For this, the local clusters are aligned by maximizing a similarity measure, specifically the cosine similarity of adjacency encoded vectors. Green et al. (2004) and Ustalov et al. (2019) produce prototypical frames, i.e., templates, and do not cluster the textual occurrences directly, while we use contextualized embeddings of the textual occurrence, which, on the other hand, increases computational complexity.

The approach by Anwar et al. (2019) performed best on the SemEval challenge on Unsupervised Lexical Frame Induction (QasemiZadeh et al., 2019), and uses separate word- and context representations from Word2Vec (Mikolov et al., 2013), and ELMo (Peters et al., 2018) as input for agglom-

[2]The code is available at https://github.com/remstef/SSC4Frames

[3]While the authors describe their approach as FIn, we agree with Yamada et al. (2023) that this is actually FId.

erative clustering. Samih and Kallmeyer (2023) extended this idea and fine-tuned contextualized embeddings with a transformer-based denoising autoencoder as the base representation for the frame-evoking elements, and also utilized clustering in a single pass. Despite achieving competitive results in benchmarks, these methods do not necessarily scale to large amounts of data due to the single-pass clustering bottleneck.

Yamada et al. (2021) introduced a two-step clustering procedure. In their experiments they show that using masked word embeddings relaxes the influence of irrelevant surface information of frame-evoking verbs and that two-step clustering improves the number of resulting frame clusters for instances of the same verb. For the two-step clustering method, first, occurrences of the same verb type are clustered, the clusters are then used for further clustering across verbs, and finally, each generated cluster is regarded as an induced frame. Similarly, Remus (2023) presented a two-step clustering approach in combination with a selection of scalable graph-clustering algorithms. They show that the methodology is able to scale to large data. Our approach uses a similar methodology, but it can make use of labeled data to improve the results.

Yamada et al. (2023) also integrated labeled data for supervised FIn where the representations of targets are fine-tuned on an FId task before applying unsupervised clustering. This is a similar setup to our approach: both use a labeled training set, to improve the quality of the clustering. However, Yamada et al.'s (2023) approach cannot be easily used for FId since the resulting clusters do not necessarily contain only examples for a specific frame, while our SSC approach keeps existing labels intact.

All these approaches use clustering for FIn. Clustering is typically used as an unsupervised method. Observations are grouped into sets regarding a given similarity measure; pre-existing labels are ignored.[4]

Semi-supervised clustering addresses this shortcoming and uses information about some observations to guide and improve the clustering process. Bair (2013); Dinler and Tural (2016); Cai et al. (2023) give overviews of the basic ideas employed in this direction. For our purposes, two aspects are important: First, the information used to guide the clustering can take different forms, like constraints or labels (Dinler and Tural, 2016). In our case, parts of the data are labeled with an incomplete label set. Second, constraints on the clusterings can be treated as hard or soft constraints (Cai et al., 2023). To be able to either label instances with a known frame, as in FId, or group them with semantically similar unlabeled instances, as in FIn, we need to ensure that clusters contain only labeled instances sharing the same label. Hence, we need to treat the labels as hard constraints that are not allowed to be violated in the resulting clustering.

We decided to use the CW (Biemann, 2006) clustering algorithm as a basis for our SSC approach, since it allows for scalability and has been successfully used for unsupervised FIn (Remus, 2023).

3. Method

In the presented approach, we treat the identification and induction of frames as a clustering problem in such a way that instances with known class labels from the training set are clustered with unseen, unlabeled instances from the test set. For this, we are using SSC. For each instance in the test set, the evoked frame label is either known from the training data or unknown, i.e., a new, unnamed frame has been discovered. The aim is to assign each target that evokes a known frame with the respective label and group the targets that evoke unknown frames into one or multiple groups, such that instances with the same frame belong to the same group. The SSC approach achieves this by allowing unlabeled instances to be grouped with labeled instances, which translates to a supervised classification, or it allows unlabeled instances to form new groups, which translates to an unsupervised clustering problem. We employ an adapted version of CW (Biemann, 2006) which guarantees that the clusters, induced by the labels of the training set, are part of the final clustering. I.e., our proposed adaptation of CW implements a hard constrained clustering with must-link and cannot-link constraints coming from the labeled data. Hence, each resulting cluster can either be labeled with an existing frame as in FId or it is a cluster of unknown but similar instances as in FIn. The new clusters are still unnamed, i.e., meaningful class labels for new frames are not generated by our method, but can be deduced post hoc, e.g., either by manual inspection of the contained examples or automatically by leveraging LLMs (Han et al., 2024).

In the following paragraphs, we present the individual steps of our method in detail.

Target identification The first step is to identify the frame-evoking elements, which are subsequently clustered. For now, we restrict ourselves to the subset of verbal LUs. This makes the method easy to apply to real-world data, since all tokens, tagged as verbs, by a part-of-speech tagger are used as targets.

[4] An alternative to these clustering-based induction methods would be approaches leveraging large-language models (LLMs) for frame induction (Torrent et al., 2024). Guo et al. (2024), however, find that LLMs perform poorly for frame induction. We leave it for future work to integrate generative LLMs into FIn.

Target representation For the clustering, we represent the targets by the first contextualized embedding of their corresponding sub-tokens from a transformer-based (masked) language model. For comparison purposes, and in accordance with recent approaches in FIn (Yamada et al., 2021; Samih and Kallmeyer, 2023) and FId (Sikos and Padó, 2019), we use BERT embeddings (Devlin et al., 2019).[5]

We follow Yamada et al. (2021) and use a weighted average between embeddings where the target is either unmasked (with weight α) or masked (with weight $1-\alpha$) in the input. α is a hyperparameter in our experiments.

Local-global clustering In order to make the SSC approach applicable to large corpora, we use a local-global clustering strategy, which is inspired by the two-step clustering idea of Yamada et al. (2021) and Remus (2023). It follows a *split-and-merge* strategy, i.e., surface form occurrences of a particular verb within its context, defined by the lemmatization, are split such that their polysemic meanings, evoking different frames, form clusters of instances (local-step), which are then merged with synonymic usages, evoking the same frame, across all verb types (global step), see Figure 1 for an illustration. This procedure is agnostic of the clustering method itself; it can thus be described as a meta-algorithm similar to Watset (Ustalov et al., 2019).

More formally, given

T, a set of frame evoking elements and their contexts, in our case the verb tokens with the sentences they appear in,

P_T, a partition of T, in our case this partition is given by the lemmas $v \in V$ of the frame evoking verb tokens,

e, a function that maps the frame evoking elements $t \in T$ to representations for a clustering, in our case $e : T \to \mathbb{R}^n$ maps the target verbs to their BERT embeddings, and

avg_e, a function that maps sets of target representations $c_i^v \in P(e(T))$ to a single representation for clustering, we use the average embedding, $avg_e : P(e(T)) \to \mathbb{R}^n, c_i^v \in P(e(T)) \mapsto \hat{x}_i^v = \frac{1}{|c_i^v|} \sum_{x \in c_i^v} x$,

the local-global clustering is defined by the following four steps:

1. *Local clustering:* Cluster the sets $e(X_v), X_v \in P_T$ individually. Here, the aim is to group the instances of a specific verb v into n_v different senses. Let the resulting clusters be c_i^v.
2. Collect all clusters from the individual local clusterings in a set $L = \bigcup_{X_v \in P_T} \{ c_1^v, \ldots, c_{n_v}^v \mid c_i^v \subseteq e(X_v) \}$.
3. Prepare the representations for the global clustering step: $L_R = \{ avg_e(c_i^v) \mid c_i^v \in L \}$.
4. *Global clustering:* Cluster L_R such that the local clusters are organized into m global clusters.

When using semi-supervised clustering, the input to the clustering algorithms does not simply consist of representations for the instances or local clusters, but of pairs of representation and a label $l \in F \cup \{\emptyset\}$, where F is the set of frame labels and $\emptyset$ is used for unlabeled instances. Please note that the labels from the instances can be used to label the local clusters. Since we use clustering algorithms with hard constraints, each local cluster contains exclusively instances with the same label or unlabeled instances, such that it is straightforward to assign this label to the local cluster itself and, as such, to the unlabeled instances. If all instances in a local cluster are unlabeled, i.e., the algorithm found a new potential frame, the local cluster is unlabeled as well, and a unique cluster ID is used. Let $\mathrm{lab} : L \to F \cup \{\emptyset\}$ be a function that assigns the labels to the local clusters.

Since local clusters might thus be assigned the same label across different verb lemmas, we test three different strategies to make use of those labels for the input of the global clustering step:

Global clustering with labels (`Label`): This is the most straightforward strategy; the representations of the local clustering are clustered using a semi-supervised clustering approach, which respects the labels of the local clusters. The input labels to the semi-supervised global clustering are simply forwarded: $L_R = \{ (avg_e(c_i^v), \mathrm{lab}(c_i^v)) \mid c_i^v \in L \}$.

Global clustering with distinct labels by merger (`Merger`): For this strategy, any two local clusters c_i^v and c_j^w with the same label are merged before applying avg_e, i.e., L will contain $avg_e(\cup\{c_i^v, c_j^w\})$ instead of $avg_e(c_i^v)$ and $avg_e(c_j^w)$. $L_R = \{ (avg_e(\cup_{c_i^v \in L, \mathrm{lab}(c_i^v)=l} c_i^v), l) \mid l \in F \} \cup \{ (avg_e(c_i^v), \emptyset) \mid c_i^v \in L, \mathrm{lab}(c_i^v) = \emptyset \}$.

Global clustering with distinct labels (`Distinct`): This strategy makes the local cluster labels unique before applying the global clustering step. We do this by combining the frame label with the verb as a partition-specific label. This prefix is removed after the global clustering, and the

[5] Also, we found more recent embedding models with a higher parameter count to encode too much sentence information in the word embeddings, and thus the performance of the method decreased eventually. We tested the NV-Embed-v2 with 7B parameters (https://huggingface.co/nvidia/NV-Embed-v2, visited October 2025), which scored among the highest at the MTEB leaderboard (https://huggingface.co/spaces/mteb/leaderboard; Enevoldsen et al., 2025) at the time of our experimentation phase.

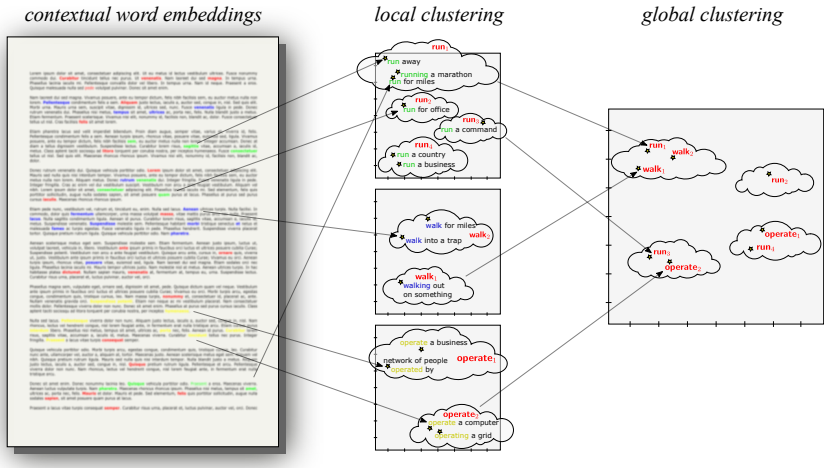

Figure 1: Schematic overview of the local-global clustering procedure.[6]

Listing 1: The generic Chinese Whispers (CW) algorithm as defined in (Biemann, 2006). Nodes with the same class label build the final clusters.

```
1 // initialize
2 forall v_i in V: class(v_i)=i;
3 // propagate
4 while changes:
5   forall v in V, randomized order:
6     class(v)=highest ranked class in the
          neighborhood of v;
```

Listing 2: Semi-supervised adaptation of CW.

```
1 // initialize
2 forall v_i in V \ V_labeled: class(v_i)=highest
      ranked class in labeled neighborhood of v_i
      if labeled neighborhood is not ∅ else i;
3 // propagate
4 while changes:
5   forall v in V \ V_labeled, randomized order:
6     class(v)=highest ranked class in the
          neighborhood of v;
```

resulting clusters are merged to obtain the final clustering. The input to the global clustering is: $L_R = \{\, (avg_e(c_i^v), (v, \text{lab}(c_i^v))) \mid c_i^v \in L \,\}$.

We implement the `Distinct` strategy to address two potential shortcomings of the two other merging strategies:

a) When directly clustering the local clusters using the frame labels, frames which are evoked by many different verbs might attract too many unlabeled local clusters. This is due to the fact that each label only appears once per verb. *b*) When merging local clusters with the same label before applying the global clustering, the labeled clusters, each cluster containing only instances from the same verb differing between the clusters, might be scattered too far across the embedding space, which leads to worse representations using the avg_e function.

In our experiments, we treat these three different strategies as a hyperparameter.

Semi-supervised CW Regarding a specific clustering algorithm, we decided to opt for a graph clustering algorithm that does not depend on specifying the number of clusters as a hyperparameter. We used CW (Biemann, 2006), which is based on a label-propagation procedure (cf. Listing 1), mainly for the following three reasons: i) CW has been developed in the context of natural language processing and has been successfully applied to FIn before, ii) CW is efficient and scalable which aligns with our goal to make the procedure applicable to large-scale corpora and iii) it is easily adaptable to a semi-supervised clustering approach as described below. Since CW expects a graph, we first compute the pairwise similarities of the target representations: nodes then represent the instances and edges are weighted by the similarity value between the representations. In our experiments, we use cosine distance. We then reduce the number of edges by applying a threshold θ to the weights: edges with a weight below θ are removed and edges with a weight higher or equal θ are kept as unweighted edges.[7]

For the SSC, we modify CW such that nodes with a given label do not update, but still propagate their label. Our modifications can be seen in Listing 2, where the changes from the original CW algorithm (cf. Listing 1) are located in Line 2 and Line 5. We introduce the set of pre-labeled nodes as $V_{labeled}$ and define the labeled neighborhood of a node to be the set of neighboring nodes $\in V_{labeled}$. Note that V still contains all nodes, labeled and unlabeled, and the default neighborhood of a node v contains labeled and unlabeled nodes. First, we change the initialization procedure in Line 2, where the default implementation assigns a unique class label per

[6] Figure taken from Remus (2023).

[7] Pruning the fully connected similarity graph results in a sparser network, such that it exhibits the small-world network property essential for CW.

node, we assign a unique new label only for nodes that are not pre-labeled and not connected to a pre-labeled node, otherwise, we treat the given label of pre-labeled instances as an already assigned and fixed cluster label. If a node is connected to one or more pre-labeled nodes, we apply CWs label propagation procedure and select the highest-ranked class in its labeled neighborhood. Second, we only update class labels for unlabeled nodes in Line 5, i.e., nodes that are not pre-labeled. With those simple but very effective changes, we can propagate class labels of existing, pre-labeled instances, thus allowing unseen instances to be classified with a known frame label, and we are also able to discover new frames, in which case, unseen instances build new, unnamed clusters. We use the modified CW as a tangible clustering implementation in both steps of the local-global clustering procedure. The pruning threshold θ is tuned as a hyperparameter for the local and the global clustering steps individually.

4. Experiments

We test our approach in two settings: frame identification and frame induction.

Frame identification is a typical supervised labeling setup. Given labeled training and development sets, the aim is to apply the frame labels to the instances in the test set. However, frame labels that do not appear in the training set are omitted from the development and test set. The respective instances are labeled with an `<unknown>` label. Hence, the task is to label the test instance either with one of the frame labels known from the training set or to label it as `<unknown>`. We report the accuracy of the labeling.

As comparison baseline we use a simple softmax or multinomial logistic regression classifier.[8] This classifier can only predict the labels from the training set. However, the standard data splits used for frame identification experiments have only a few instances with unknown labels in the test set (cf. Table 1).

For the BFN data, we also report results for the established tool OpenSesame (Swayamdipta et al., 2017). Please note that our results differ from the results reported in the paper. One reason is that we limited ourselves to verbal targets. Furthermore, during our experiments we noted that OpenSesame uses the whole BFN, including development and test set, to learn backup labels. Thus it applies labels to test set instances that do not appear in the training set based on their appearance in the development and test sets. We treated instances labeled with such labels as false predictions.

The results reported for current state-of-the-art methods are higher, however, the target representations used in current approaches employ frame representations as well. Therefore they are harder to adapt to combined frame identification and frame induction. We leave this for possible future work.

Frame induction is a typical unsupervised clustering setup. Given the instances in the train and development or the train and test sets respectively, we evaluate the clustering of the development/test set compared to the clustering induced by the frame labels. For our semi-supervised clustering approach, we use the labels in the training set, while unsupervised approaches only utilize the unlabeled instances from the training set. We report, B^3-Recall (B^3R), B^3-Precision (B^3P) and their harmonic mean (B^3F1) (Bagga and Baldwin, 1998).[9]

As comparison methods, we report the results of the approach by Yamada et al. (2021), using both hierarchical agglomerative clustering with group average linkage (GA) and X-means (Pelleg and Moore, 2000) for the local clustering step. We follow the hyperparameter tuning approach described by Yamada et al. (2021).

We also compare the impact of the semi-supervised approach by reporting the results of using the standard unsupervised CW.

Local-global clustering is introduced in our setup in order to allow the scalability of the approach. However, we also want to test, how this approach has an effect on the performance both in FId and FIn. Therefore, we also report the results of a one-step approach, i.e., simply using CW for clustering the data. In the frame identification setup, we only use semi-supervised CW since the results need to be labeled. In the frame induction method, we report both the results of unsupervised and semi-supervised CW.

As a simple baseline for the local clustering step, we also introduce the one-cluster-per-verb strategy, where all unlabeled instances for a given verb are grouped together in one cluster. This makes use of the fact that many verbs in the datasets are not ambiguous regarding evoked frames.

We test our approach on English and German using the BFN, version 1.7, and SALSA, version 2.0, datasets respectively. We treat only verbs as frame evoking elements and therefore discard all non-verbal LUs.

For the BFN, we use the same data split as Swayamdipta et al. (2017) and use an uncased

[8] We used the LogisticRegression model from scikit-learn (Pedregosa et al., 2011) with lbfgs solver. We optimized C and compared using the embeddings as is or applying a standardization using StandardScaler from scikit-learn on the development set.

[9] Yamada et al. (2021) also report purity, inverse purity and their harmonic mean for their experiments. Since these metrics highly correlate with the B^3 metrics in our experiments, we omitted them.

BERT model for the target representations.[10]

For SALSA, the data is split into training, development and test sets following Botschen et al. (2018). We only use verbal targets not annotated with proto-frames. Proto-frames are a concept used for the annotation of the SALSA dataset in cases where the evoked frame is unknown with regard to the label set. For a given verb, such instances are grouped together if they evoke the same unknown frame. However, it is not annotated whether different verbs can evoke the same proto-frame. Therefore, we cannot use them for evaluating frame induction. For the target representations, we use a cased BERT model for German.[11]

Statistics for both datasets are shown in Table 1. One obvious difference between the datasets is that BFN has fewer examples per verb and per frame. This can be seen with the test sets: while comparable in size regarding instances, BFN has more different verbs and frames. Apart from language specific differences, this might also be due to the creation of the datasets: the instances in SALSA come from full-text annotations while the instances in the BFN dataset have been collected in order to illustrate the different frames. This difference has also an impact on the experiment results since the proportion of verbs that are labeled with multiple frames is higher in the SALSA dataset: 203 out of the 418 verbs are ambiguous compared to 248 out of the 1285 verbs in BFN.

5. Evaluation

In this section, we present the best hyperparameters from our tuning on the development set and the results of the best settings on the test set.

5.1. Hyperparameter tuning

For all clustering algorithms, we do hyperparameter tuning by clustering training and development instances and optimizing for the given metric, i.e., accuracy for frame identification and B^3F1 for frame induction, on the development set.

Whenever using BERT embeddings, we take the weighted average between masked and unmasked embeddings and tune the weight α using a grid search for values between 0, using only the masked embedding, and 1, using only the unmasked embedding, with 0.1 steps. Most of the values are either 0.7, 0.8, 0.9 or 1 for all tasks, algorithms and both datasets. I.e., the unmasked embeddings are favored. One notable exception are the best α values for our main approach, i.e., two-step semi-supervised CW. They are 0 for FIn and FId on SALSA, and 0.2 for FIn respectively 0.5 for FId on BFN.

For the approach by Yamada et al. (2021), the values are either 0.9 or 1. I.e., only the unmasked embeddings are used or at least favored. This differs from the results reported by Yamada et al. (2021).

For our local-global SSC, we take the merging strategy working best on the development set. For most of the settings and and datasets, the best merging strategy is `Label`. The only exceptions are FIn on BFN using the baseline one-cluster-per-verb strategy for the local clustering step, FIn on SALSA with two-step CW and FId on SALSA using the one-cluster-per-verb strategy. In these settings, the best merging strategy is `Distinct`.

For CW, we tune the distance threshold θ used to prune the graph. The best values are between 0.5 and 0.9 for both local and global clusterings.

5.2. Test results

For testing, training and test instances are clustered. The difference between unsupervised and semi-supervised clustering algorithms is that the semi-supervised algorithm gets the labeled training instances as input.

For algorithms that have non-deterministic elements, we did five runs with different initializations. With CW, the clustering results can depend on the order in which the nodes are updated (cf. Line 5 in Listing 1 and 2). For local-global clusterings, we did five runs for each, the local and the global step, i.e., 25 runs in total. In these cases, we report the mean with standard deviation. The results show low standard deviations overall.

Frame induction The results (cf. Table 2) show that our local-global clustering approach with unsupervised CW as clustering algorithm is by itself a strong starting point. Regarding B^3F1 it outperforms the approach of Yamada et al. (2021) for both datasets. Only in terms of B^3R the approach of Yamada et al. (2021) leads to better results.

Adding supervision with our semi-supervised adaptation of CW improves the results further, leading to better results than Yamada et al. (2021) regarding all metrics. The training labels especially improve B^3R. For the BFN dataset, the unsupervised CW even leads to slightly better results regarding B^3P. However, B^3F1 is improved by the training labels for both datasets.

While we use the local-global approach especially for scalability, the experiment results also show improved results compared with the one-step clustering.

[10] https://huggingface.co/google-bert/bert-base-uncased (visited October 2025).

[11] https://huggingface.co/google-bert/bert-base-german-cased (visited October 2025).

	BFN				SALSA			
	total	train	dev	test	total	train	dev	test
instances	8329	5739	686	1904	12536	8730	1883	1923
verbs	1285	1072	294	547	418	386	273	272
frames	471	428	190	314	237	226	166	177
unknown frames								
instances	75		10	65	14		6	8
verbs	52		10	44	12		6	8
frames	43		10	39	11		6	8

Table 1: Dataset statistics.

Method	BFN			SALSA		
	B^3P	B^3R	B^3F1	B^3P	B^3R	B^3F1
Yamada et al. (2021)						
GA	0.53	0.70	0.60	0.45	0.58	0.51
X-means	0.53	0.68	0.60	0.39	0.59	0.47
One-Step						
CW	0.75 ± 0.01	0.44 ± 0.00	0.55 ± 0.00	0.37 ± 0.01	0.63 ± 0.01	0.47 ± 0.01
CW-semi	0.80 ± 0.01	0.46 ± 0.00	0.59 ± 0.00	0.84 ± 0.00	0.45 ± 0.00	0.58 ± 0.00
One-cluster-per-verb						
CW	0.84 ± 0.00	0.64 ± 0.00	0.72 ± 0.00	0.58 ± 0.01	0.57 ± 0.01	0.57 ± 0.01
CW-semi	0.81 ± 0.00	0.67 ± 0.00	0.73 ± 0.00	0.70 ± 0.01	0.69 ± 0.01	0.69 ± 0.00
CW	0.87 ± 0.00	0.63 ± 0.00	0.73 ± 0.00	0.72 ± 0.00	0.50 ± 0.00	0.59 ± 0.00
CW-semi	0.81 ± 0.00	0.77 ± 0.00	0.79 ± 0.00	0.79 ± 0.00	0.78 ± 0.00	0.78 ± 0.00

Table 2: Test results for frame induction.

Method	BFN	SALSA
Softmax	0.61	0.82
OpenSesame	0.69 ± 0.01	
One-Step CW-semi	0.33 ± 0.01	0.57 ± 0.00
1cpv + CW-semi	0.40 ± 0.02	0.64 ± 0.01
CW-semi	0.69 ± 0.00	0.82 ± 0.00

Table 3: Test results for frame identification (Accuracy). 1cpv denotes the one-cluster-per-verb strategy for the local clustering step.

Frame identification Regarding FId our SSC approach leads to the best results for both datasets (cf. Table 3). They are on a par with the results from the softmax classifier and OpenSesame. Here, one-step clustering has the lowest accuracy. The better results on the SALSA dataset can be explained by the fact that the dataset contains fewer frames and more instances, i.e., more examples per frame (cf. Section 4 and Table 1).

6. Conclusion and further work

In this paper, we introduced SSC for joint FId and FIn in order to overcome a) the problem of FId to deal with unknown frames in real-world corpora when approached as a supervised multi-class labeling problem and b) the problem that labeled data is not used in FIn when approached as an unsupervised clustering task.

As a concrete implementation, we evaluated an adapted version of the graph clustering algorithm CW in a local-global clustering approach using an average of masked and unmasked BERT-embeddings as representation for the FEEs. In future work, we want to further explore our semi-supervised adaptation of CW in comparison to other SSC algorithms.

We have shown that SSC improves the performance of CW in a typical FIn setting. Furthermore, our approach also outperforms current state-of-the-art FIn methods. Regarding FId, our approach reaches good results, comparable with those of the established tool OpenSesame.

Due to the scalability of CW and the local-global approach, our method can be applied to large corpora. In future work, we would like to evaluate the scalability and the performance on a real-world dataset. We would like to explore ways how SSC of such a dataset like Wikipedia can be efficiently used for the expansion of a FrameNet.

Furthermore, we would like to evaluate different representations for the LUs. Recent work has

shown that BERT-embeddings can be improved by fine-tuning. This has been done supervised on data labeled with frames (cf. Sikos and Padó, 2019 for FId and unsupervised (Samih and Kallmeyer, 2023). Work in FId also utilize frame definitions (cf. Section 2) for representation of labeled data. Even simply using embeddings from more recent LLMs might improve the performance, however, in preliminary experiments with current state-of-the-art embeddings we found decreased performance (cf. Footnote 5). We would like to investigate the representation of targets in more detail in future work.

Finally, we would like to investigate how our approach works with non-verbal LUs. This is an interesting extension, since existing FrameNets focus heavily on verbal LUs. Hence, the extension of FrameNets with non-verbal LUs is an important future step for FrameNet development.

7. Bibliographical 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, MN, USA.

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 79–85, Montréal, QC, Canada.

Eric Bair. 2013. Semi-supervised clustering methods. *WIREs Computational Statistics*, 5(5):349–361.

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*, pages 86–90, Montréal, QC, Canada.

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, NY, USA.

Teresa Botschen, Iryna Gurevych, Jan-Christoph Klie, Hatem Mousselly-Sergieh, and Stefan Roth. 2018. Multimodal frame identification with multilingual evaluation. 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 1481–1491, New Orleans, LA, USA.

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. 2006. The SALSA corpus: a German corpus resource for lexical semantics. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 969–974, Genoa, Italy.

Jianghui Cai, Jing Hao, Haifeng Yang, Xujun Zhao, and Yuqing Yang. 2023. A review on semi-supervised clustering. *Information Sciences*, 632:164–200.

Dipanjan Das, Nathan Schneider, Desai Chen, and Noah A. Smith. 2010. Probabilistic frame-semantic parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 948–956, Los Angeles, CA, USA.

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, MN, USA.

Derya Dinler and Mustafa Kemal Tural. 2016. A survey of constrained clustering. In M. Emre Celebi and Kemal Aydin, editors, *Unsupervised Learning Algorithms*, pages 207–235. Springer International Publishing, Cham.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Veysel Çağatan, Akash Kundu, Martin Bernstorff, Shitao Xiao, Akshita Sukhlecha, Bhavish Pahwa, Rafał Poświata, Kranthi Kiran GV, Shawon Ashraf, Daniel Auras, Björn Plüster, Jan Philipp Harries, Loïc Magne, Isabelle Mohr, Dawei Zhu, Hippolyte Gisserot-Boukhlef, Tom Aarsen, Jan Kostkan, Konrad Wojtasik, Taemin Lee, Marek

Suppa, Crystina Zhang, Roberta Rocca, Mohammed Hamdy, Andrianos Michail, John Yang, Manuel Faysse, Aleksei Vatolin, Nandan Thakur, Manan Dey, Dipam Vasani, Pranjal A Chitale, Simone Tedeschi, Nguyen Tai, Artem Snegirev, Mariya Hendriksen, Michael Günther, Mengzhou Xia, Weijia Shi, Xing Han Lù, Jordan Clive, Gayatri K, Maksimova Anna, Silvan Wehrli, Maria Tikhonova, Henil Shalin Panchal, Aleksandr Abramov, Malte Ostendorff, Zheng Liu, Simon Clematide, Lester James Validad Miranda, Alena Fenogenova, Guangyu Song, Ruqiya Bin Safi, Wen-Ding Li, Alessia Borghini, Federico Cassano, Lasse Hansen, Sara Hooker, Chenghao Xiao, Vaibhav Adlakha, Orion Weller, Siva Reddy, and Niklas Muennighoff. 2025. MMTEB: Massive multilingual text embedding benchmark. In *Proceedings of the Thirteenth International Conference on Learning Representations*, Singapore.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

Charles J. Fillmore. 1982. Frame semantics. In Linguistic Society of Korea, editor, *Linguistics in the Morning Calm. Selected Papers from SICOL-1981*, pages 111–137. Hanshin Publishing Company, Seoul, South Korea.

Rebecca Green, Bonnie J. Dorr, and Philip Resnik. 2004. Inducing frame semantic verb classes from WordNet and LDOCE. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 375–382, Barcelona, Spain.

Shaoru Guo, Yubo Chen, Kang Liu, Ru Li, and Jun Zhao. 2024. NutFrame: Frame-based conceptual structure induction with LLMs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12330–12335, Torino, Italia.

Yi Han, Ryohei Sasano, and Koichi Takeda. 2024. Definition generation for automatically induced semantic frame. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11112–11118, Bangkok, Thailand.

Silvana Hartmann, Ilia Kuznetsov, Teresa Martin, and Iryna Gurevych. 2017. Out-of-domain FrameNet semantic role labeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 471–482, Valencia, Spain.

Karl Moritz Hermann, Dipanjan Das, Jason Weston, and Kuzman Ganchev. 2014. Semantic frame identification with distributed word representations. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1448–1458, Baltimore, MD, USA.

Tianyu Jiang and Ellen Riloff. 2021. Exploiting definitions for frame identification. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2429–2434, Online.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the 1st International Conference on Learning Representations*, Scottsdale, AZ, USA.

Alexis Palmer and Caroline Sporleder. 2010. Evaluating FrameNet-style semantic parsing: the role of coverage gaps in FrameNet. In *Coling 2010: Posters*, pages 928–936, Beijing, China.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Dan Pelleg and Andrew W. Moore. 2000. X-means: Extending k-means with efficient estimation of the number of clusters. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, page 727–734, San Francisco, CA, USA.

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*, pages 2227–2237, New Orleans, LA, USA.

Alexander Popov and Jennifer Sikos. 2019. Graph embeddings for frame identification. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 939–948, Varna, Bulgaria.

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*, pages 16–30, Minneapolis, MN, USA.

Ines Rehbein, Josef Ruppenhofer, Caroline Sporleder, and Manfred Pinkal. 2012. Adding nominal spice to SALSA - frame-semantic annotation of German nouns and verbs. In *11th Conference on Natural Language Processing, KONVENS 2012, Empirical Methods in Natural Language Processing*, volume 5 of *Scientific series of the ÖGAI*, pages 89–97, Vienna, Austria.

Steffen Remus. 2023. *Domain Defining Context: On Domain-Dependent Corpus Expansion and Contextualized Semantic Structuring*. Ph.D. thesis, Staats-und Universitätsbibliothek Hamburg Carl von Ossietzky.

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.

Younes Samih and Laura Kallmeyer. 2023. Unsupervised semantic frame induction revisited. In *Proceedings of the 15th International Conference on Computational Semantics*, pages 89–93, Nancy, France.

Jennifer Sikos and Sebastian Padó. 2019. Frame identification as categorization: Exemplars vs prototypes in embeddingland. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 295–306, Gothenburg, Sweden.

Xuefeng Su, Ru Li, Xiaoli Li, Jeff Z. Pan, Hu Zhang, Qinghua Chai, and Xiaoqi Han. 2021. A knowledge-guided framework for frame identification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5230–5240, Online.

Swabha Swayamdipta, Sam Thomson, Chris Dyer, and Noah A. Smith. 2017. Frame-semantic parsing with softmax-margin segmental RNNs and a syntactic scaffold. *CoRR*, abs/1706.09528.

Fabio Tamburini. 2022. Combining ELECTRA and adaptive graph encoding for frame identification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1671–1679, Marseille, France.

Sang-Sang Tan and Jin-Cheon Na. 2019. Positional attention-based frame identification with BERT: A deep learning approach to target disambiguation and semantic frame selection. *CoRR*, abs/1910.14549.

Tiago Timponi Torrent, Thomas Hoffmann, Arthur Lorenzi Almeida, and Mark Turner. 2024. *Copilots for Linguists: AI, Constructions, and Frames*. Elements in Construction Grammar. Cambridge University Press.

Shogo Tsujimoto, Kosuke Yamada, and Ryohei Sasano. 2025. Semantic frame induction from a real-world corpus. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 991–997, Vienna, Austria.

Dmitry Ustalov, Alexander Panchenko, Chris Biemann, and Simone Paolo Ponzetto. 2019. Watset: Local-global graph clustering with applications in sense and frame induction. *Computational Linguistics*, 45(3):423–479.

Kosuke Yamada, Ryohei Sasano, and Koichi Takeda. 2021. 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.

Kosuke Yamada, Ryohei Sasano, and Koichi Takeda. 2023. Semantic frame induction with deep metric learning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1833–1845, Dubrovnik, Croatia.

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*, pages 1247–1256, Copenhagen, Denmark.

Zheng Xin Yong and Tiago Timponi Torrent. 2020. Semi-supervised deep embedded clustering with anomaly detection for semantic frame induction. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3509–3519, Marseille, France.