SemEval-2019 Task 2: Unsupervised Lexical Frame Induction

Behrang QasemiZadeh
SFB991, Germany
zadeh@phil.hhu.de

Miriam R. L. Petruck
ICSI, US
miriamp@icsi.berkeley.edu

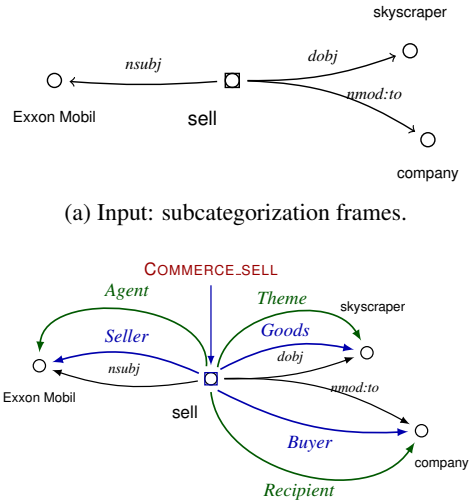
Regina Stodden
HHUD, Germany
stodden@phil.hhu.de

Laura Kallmeyer
HHUD, SFB991, Germany
kallmeyer@phil.hhu.de

Marie Candito
Paris Diderot University - CNRS, France
marie.candito@linguist.univ-paris-diderot.fr

Abstract

This paper presents Unsupervised Lexical Frame Induction, Task 2 of the International Workshop on Semantic Evaluation in 2019. Given a set of prespecified syntactic forms in context, the task requires that verbs and their arguments be clustered to resemble semantic frame structures. Results are useful in identifying polysemous words, i.e., those whose frame structures are not easily distinguished, as well as discerning semantic relations of the arguments. Evaluation of unsupervised frame induction methods fell into two tracks: Task A) *Verb Clustering* based on FrameNet 1.7; and B) *Argument Clustering*, with B.1) based on FrameNet's core frame elements, and B.2) on VerbNet 3.2 semantic roles. The shared task attracted nine teams, of whom three reported promising results. This paper describes the task and its data, reports on methods and resources that these systems used, and offers a comparison to human annotation.



(a) Input: subcategorization frames.

(b) Output: Semantic Frame Tagging using labels learned by Unsupervised methods.

Figure 1: Given semantically unlabeled structures (1a), annotate the input with semantic information learned via unsupervised methods (1b).

1 Introduction

SemEval 2019 Task 2 focused on the unsupervised semantic labeling of a set of prespecified (semantically) unlabeled structures (Figure 1). Unsupervised learning methods analyze these structures (Figure 1a) to augment them with semantic labels (Figure 1b). The shape of the manually labeled input frames is constrained to an *acyclic* connected tree of lexical items (words and multi-word units) of maximum depth 1, where just one root governs several arguments. The task used Berkeley FrameNet (FN) (Ruppenhofer et al., 2016) and Q. Zadeh and Petruck (2019), guidelines for this task, to determine the arguments and label them with semantic information.

We compared the proposed system results for unsupervised semantic tagging with that of human annotated (or, gold-standard) data in three different subtasks (Figure 2). To evaluate the systems, we computed distributional similarities between their generated unsupervised labeled data and human annotated reference data. For computing similarities we used general purpose numeral methods of text clustering, in particular BCUBED F-SCORE (Bagga and Baldwin, 1998) as the single figure of merit to rank the systems.

The most important result of the shared task is the creation of a benchmark for a future complex task. This benchmark includes a moderately sized, manually annotated set of frames, where only the verbs of each were included, along with their *core frame elements* (which uniquely define a frame as Ruppenhofer et al. describe). To complement FN's core frame elements that have highly specific meanings, the benchmark also includes the annotated argument structures of the verbs based on the generic semantic roles proposed for verb classes in VerbNet 3.2 (Kipper et al., 2000; Palmer et al., 2017). The benchmark comes with simplified annotation guidelines and a modular annotation sys-

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 16–30
Minneapolis, Minnesota, USA, June 6–7, 2019.

tem with browsing and editing capabilities.[1] Complementing the benchmarking are several state-of-the-art competing baselines, from the participants, that serve as a point of departure for improvements in the future.[2]

The rest of this paper is organized as follows: Section 2 contextualizes this task; Section 3 offers a detailed task-description; Section 4 describes the data; Section 5 introduces the evaluation metrics and baselines; Section 6 characterizes the participating systems and unsupervised methods that participants used; Section 7 provides evaluation scores and additional insight about the data; and Section 8 presents concluding remarks.

2 Background

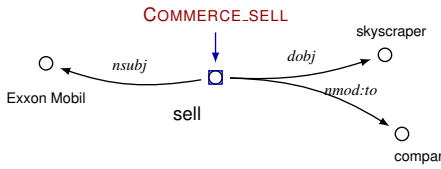
Frame Semantics (Fillmore, 1976) and other theories (Gamerschlag et al., 2014) that adopt typed feature structures for representing knowledge and linguistic structures have developed in parallel over several decades in theoretical linguistic studies about the syntax–semantics interface, as well as in empirical corpus-driven applications in natural language processing. Building repositories of (lexical) semantic frames is a core component in all of these efforts. In formal studies, lexical semantic frame knowledge bases instantiate foundational theories with tangible examples, e.g., to provide supporting evidence for the theory. Practically, frame semantic repositories play a pivotal role in natural language understanding and semantic parsing, both as inspiration for a representation format and for training data-driven machine learning systems, which is required for tasks such as information extraction, question-answering, text summarization, among others.

However, manually developing frame semantic databases and annotating corpus-derived illustrative examples to support analyses of frames are resource-intensive tasks. The most well-known frame semantic (lexical) resource is FrameNet (Ruppenhofer et al., 2016), which only covers a (relatively) small set of the vocabulary of contemporary English. While NLP research has integrated FrameNet data into semantic parsing, e.g., Swayamdipta et al. (2018), 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, a common challenge in semantic parsing (Hartmann et al., 2017).

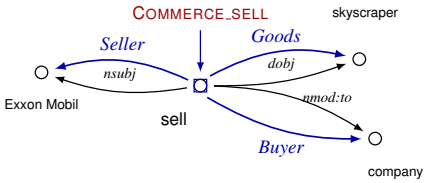
Unsupervised frame induction methods can serve as an assistive semantic analytic tool, to build language resources and facilitate linguistic studies. Since the focus is usually to build language resources, most systems (Pennacchiotti et al. (2008); Green et al. (2004)) have used a lexical semantic resource like WordNet (Miller, 1995) to extend coverage of a resource like FrameNet. Some methods, e.g., Modi et al. (2012) and Kallmeyer et al. (2018), tried to extract FrameNet-like resources automatically without additional semantic information. Others (Ustalov et al. (2018); Materna (2012)) addressed frame induction only for verbs with two arguments.

Lastly, unsupervised frame induction methods can also facilitate linguistic investigations by capturing information about the reciprocal relationships between statistical features and linguistic or extra-linguistic observations (e.g., Reisinger et al. (2015)). This task aimed to benchmark a class of such unsupervised frame induction methods.

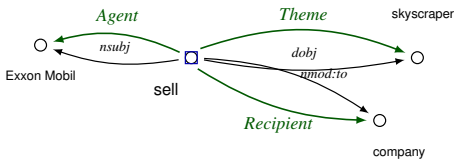
3 Task Description



(a) **Task A** - Identifying Semantic Frames: Unsupervised learned labels evaluated against FN's lexical units



(b) **Task B.1** - Full Frame Semantic Tagging: Unsupervised labels evaluated against FN's frames



(c) **Task B.2** – Case Role Labeling: Unsupervised labels evaluated against generic semantic roles (VerbNet)

Figure 2: Subtasks of SemEval 2019 Task 2.

[1] http://sfa.phil.hhu.de/.

[2] See https://competitions.codalab.org/competitions/19159 for accessing the task's language resources, tools, and further technical details.

The ambitious goal of this task was the unsupervised induction of frame semantic structures from tokenized and morphosyntacally labeled text corpora. We sought to achieve this goal by building an evaluation benchmark for three tasks. **Task A** dealt with unsupervised labeling of verb lemmas with their frame meaning. **Task B** involved unsupervised argument role labeling, where **B.1** benchmarked unsupervised labeling of frame-specific frame elements (FEs) based on FN, and **B.2** benchmarked unsupervised role labeling of arguments in Case Grammar terms (Fillmore, 1968) and against a set of generic semantic roles, taken primarily from VerbNet.

The task was unsupervised in that it forbade the use of any explicit semantic annotation (only permitting morphosyntactic annotation). Instead, we encouraged the use of unsupervised representation learning methods (e.g., word embeddings, brown clusters) to obtain semantic information. Hence, systems learn and assign semantic labels to test records without appealing to any explicit training labels. For development purposes, developers received a small labeled development set.

3.1 Task A: Clustering Verbs

The goal of this task was to identify verbs that evoke the same frame. The task involved labeling verb uses in context to resemble their categorization based on Frame Semantics (Figure 2a). Here, we used FN 1.7 as the reference for frame definitions. Hence, the task constituted the unsupervised induction of FN's lexical units, where a lexical unit (LU) is a pairing of a lemma and a frame. For example, we expected that the LUs *auction*.v, *retail*.v, *sell*.v, etc., which evoke the typed situation of COMMERCE_SELL, be labeled with the same unsupervised tag.[3]

The task resembles word sense induction in that it assigns a class (or sense) label to a verb. In word sense induction (WSI), labels are determined and evaluated on word forms (lemma + part-of-speech e.g., *sell.v* or *auction.n*). WSI evaluations assume that the inventory of senses (set S_is) for different word forms f is devised independently. For instance, assuming f_1 is labeled with the set of senses S_1 and f_2 with S_2, then $S_1 \cap S_2 \neq \phi$ only if $f_1 = f_2$; and, if $f_1 \neq f_2$ then $S_1 \cap S_2 = \phi$ (as in other SemEval benchmarks, including Agirre and Soroa (2007); Manandhar et al. (2010); Jurgens and Klapaftis (2013); Navigli and Vannella (2013)). For instance, in WSI evaluations based on OntoNotes (Hovy et al., 2006), six different labels from S_{sell} are assigned to the lemma $sell$.v, and one label s' is assigned to $auction$.v, knowing that $s' \notin S_{sell}$. Typically, lexical semantic relationships among members of S_is (e.g., synonymy, antonymy) are then analyzed independently of WSI (e.g., Lenci and Benotto (2012); Girju et al. (2007); McCarthy and Navigli (2007)). In contrast, this task assumes that the sense inventory is defined independent of word forms.

This task involves uncovering mapping between word forms f and members of S such that different word forms (i.e., $f_i \neq f_j$) can be mapped to the same meaning (label), and the same meaning (label) can be mapped to several word forms. We defined S with respect to FrameNet and assumed that its typed-situation frames are units of meaning. So, COMMERCE_SELL captures the meaning associated with both $sell$.v and $auction$.v., as well as other selling-related words. Hence, in some sense, Task A goes beyond the ordinary WSI task as it also demands identifying (unspecified) lexical semantic relationships between verbs.

3.2 Task B.1: Unsupervised Frame Semantic Argument Labeling

Taking the frames as primary and defining roles relative to each frame, the aim of Task B.1 was to cluster prespecified verb-headed argument structures according to the principles of Frame Semantics, where FrameNet served as the reference for evaluation. This task amounted to unsupervised labeling of frames and core FEs (Figure 2b). Because FrameNet defines FEs frame-specifically, Task B.1 entails Task A.

Given a set of semantically-unlabelled arguments as input (e.g., Figure 1a), the root nodes (i.e., verbs) are clustered and assigned to a set of unsupervised frame labels π_i ($1 \leq i \leq n$, where n is the number of latent frames). Then, the arguments are labeled with semantic role labels (FEs) interpreted locally given the frame. That is, for any pair of π_x and π_y, the set of assigned roles R_x to arguments under π_x are assumed to be independent from R_y labels for π_y ($R_x \cap R_y = \phi$).

3.3 Task B.2: Unsupervised Case Role Labeling

We defined Subtask B.2 in parallel to Subtask B.1 and involved an idea from Case Grammar. The ar-

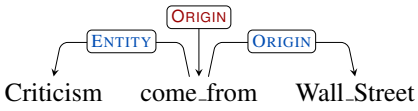
[3] Dark red small caps indicate FN frames.

guments of a verb in a set of prespecified subcategorization frames were clustered according to a common set of generic semantic roles (Figure 2c). Here, the task assumed that semantic roles are universal and generic (e.g., Agent, Patient). Their configuration determines the argument structure of verb-headed phrases. We evaluated this unsupervised labeling of arguments with semantic roles independently of the class, sense, and word form of a verb. We compared the role labels against a set of semantic roles from VerbNet 3.2 (Kipper et al., 2000). Given a verb instance, no guarantee exists that input argument structures for B.2 and B.1 would be the same.

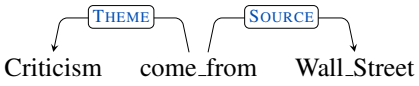
4 Evaluation Dataset

The dataset consists of manual annotations for verb-headed frame structures anchored in tokenized sentences. These frame structures were manually annotated using the guidelines for this task (Q. Zadeh and Petruck, 2019). For example, as already illustrated, the verb *come_from*.v is annotated in terms of FN's ORIGIN frame and its *core* FEs, as Example 1 shows.

(1) Criticism of futures COMES FROM Wall Street.



Also, using the set of 32 generic semantic role labels in VerbNet 3.2 and two additional roles, COGNIZER and CONTENT, we annotated arguments of the verb as the following graphic shows.



We assumed unique identifiers for sentences, e.g., #s1 for Example 1. The evaluation record for *come_from*.v (Task A) appears below, where #s1 4_5 specifies the position of the verb in the sentence (Example 1).

A [#s1 4_5 come_from.ORIGIN]

Similarly, for Task B.1 and Task B.2, respectively, the evaluation records are as follows here.

B.1 [#s1 4_5 come_from.ORIGIN Criticism-:-1-:-ENTITY Wall_Street-:-6_7-:-ORIGIN]

B.2 [#s1 4_5 come_from.NA Criticism-:-1-:-THEME Wall_Street-:-6_7-:-SOURCE]

We stripped off the manually asserted labels from the records and passed them to systems for assigning unsupervised labels. Evidently, later a scorer program (Section 5) compared system-generated labels with the manually assigned labels.

4.1 Data Sampling

We sampled data from the *Wall Street Journal* (WSJ) corpus of the Penn Treebank. Kallmeyer et al. (2018) provided frame annotations similar to those in this task for a portion of WSJ sentences, using SemLink (Bonial et al., 2013) and EngVallex (Cinková et al., 2014) to generate frame semantic annotations semi-automatically. That work was based on FrameNet and the Prague Dependency Treebank (PSD) (Hajič et al., 2012) from the Broad-coverage Semantic Dependency resource (Oepen et al., 2016). We started by annotating a portion of the records in Kallmeyer et al. (2018), and later deviated from this subset to create a more representative sample of the overall diversity and distribution of verbs in the WSJ corpus using a stratified random sampling method.

4.2 Guidelines

The annotation guidelines for this task were slightly different from those of FrameNet and various semantic dependency treebanks. In contrast to FN, which annotates a full span of text as an argument filler, or PropBank, which annotates syntactic constituents of arguments of verbs (Palmer et al., 2005), we identified the text spans and only annotated a single word or a multi-word unit (MWU), i.e., the semantic head of the span, like annotations in Oepen et al. (2016) and Abstract Meaning Representation (Banarescu et al., 2013). To illustrate, in Example 1, FN would annotate *Criticism of futures* as filling the FE ENTITY. We only annotated *Criticism*, understanding it as the LU that evokes JUDGMENT_COMMUNICATION, which in turn represents the meaning of the whole text span. Thus, we assumed that another frame f_a fills an argument of a frame. We annotated only the main content word(s) that evoke(s) f_a; these main words are the *semantic heads*.[4]

Multi-word unit semantic heads (e.g., named entities, word form combinations) are annotated as if a single word form, such as *Wall Street* (# 1), excluding modifiers. In contrast to semantic depen-

[4]The annotation guidelines (Q. Zadeh and Petruck, 2019) discuss decisions about marking semantic heads and the complex situations resulting from it for argument annotation.

dency structures (e.g., DELPH-IN MRS-Derived Semantic Dependencies, Enju PredicateArgument Structures, and Tectogramatical Representation in PSD (Oepen et al., 2016)), we did not commit to the underlying syntactic structure of the sentence since we were not obliged to relabel only syntactic structures. Rather, we annotated words and MWUs if the frame analysis permitted doing so.[5]

4.3 Annotation Procedure

We annotated the data in a modular manner and in a semi-controlled environment using an annotation system developed for this purpose. The procedure consisted of four steps: 1) Reading and Comprehension; 2) Choosing a Frame; 3) Annotating Arguments; and 4) Rating, Commenting, or Revising. We tracked and logged all changes in the data as well as annotator interaction with the annotation system upon starting to annotate. The tool measured the time that annotators spent on each record and each annotation step, as well as how annotators moved between steps.

In Step 1, annotators viewed a sentence with one highlighted verb, as in Example 2.

(2) Criticism of futures **COMES** from Wall Street.

The goal of this step was understanding the meaning of the verb and its semantic function, and identifying semantic heads of arguments and their associated words or MWUs. To continue, an annotator must confirm the understanding of the verb's meaning of the verb, and can identify its semantic arguments. Without confirmation, an annotator would terminate the annotation process for that input sentence and go to the next one.

If confirmed, Step 2 required the annotator to choose the frame that the verb evoked. This step may have included annotating multi-word phrasal verbs, e.g., COMES+FROM (Example 2). The annotation system assisted by providing a list of likely frames for the verb, including a LU lookup function (as in FN), an extended set of LUs derived via statistical methods, and previously logged annotations. After reviewing the definitions of the proposed frames, annotators chose one, or annotated the verb form with a different existing FN frame. Otherwise, the annotator terminated the process and the record moved to the list of "skipped items".

The annotation of arguments, Step 3, required that annotators label the core FEs of the chosen frame by first identifying their semantic head, which first may have required marking MWUs, e.g., Wall+Street in Example 3, below.

(3) Criticism of futures **comes from** Wall Street.

The tool lists the core FEs and their definitions, and checks the integrity of record annotations to ensure that each core FE is annotated only once. In parallel, annotators add the verb's subcategorization frame and its semantic role. We did not annotate null instantiated FEs (but FN does). During step 3, annotators could go back to the previous step and change their choice of frame type.

For Step 4, annotators rated their annotation, stating their opinion on how well the annotated instance fit FrameNet's definition and how it compared to other annotated instances. In a sense, annotators measured their confidence in the assigned labels. They did so by selecting a number on a scale from 1 to 5, with 1 not confident at all and 5 the most confident, i.e., the annotation fit perfectly to the chosen FrameNet frame, its definition, and examples. Annotators had the option to add free text comments on each record.

The annotation procedure was rarely straightforward. Given the interdependence of Steps 2 and 3, annotators usually moved back and forth between them. In Step 2 an annotator might believe that a target verb did not belong in any existing FN frame. Likewise, annotators could terminate the annotation process even upon reaching the last step.

4.3.1 Quality Control

At least two annotators verified all annotation used in the evaluation. A main annotator annotated all records in the dataset; two other annotators verified or disputed those annotations. If annotators could not reach an agreement, we removed the record from the SemEval dataset.

A full analysis of annotator disagreement goes beyond the scope of this work. While the source of annotator disagreement may seem trivial and simple (e.g., only one annotator understood the sentence correctly), we believe that some sentences may have more than one interpretation, all of which are plausible. Like the disagreement resulting from incorrect frame assignment, deciding what frame a verb evokes may be challenging; and resolving the dilemma is not always simple. Choosing between two related frames (e.g.,

[5] Q. Zadeh and Petruck describe the issues in detail.

BUILDING vs. INTENTIONALLY_CREATE, related via Inheritance in FN), or identifying metaphorical and non-metaphorical uses of a verb requires subtle and sophisticated understanding of the semantics of the language, and of Frame Semantics. At times, disagreements pointed to more complex linguistic issues that remain in debate, e.g., choosing the semantic head of a syntactically complex argument, treating quantifiers, conjunctions, etc.

4.4 Summary statistics

Table 1 shows a statistical summary of the annotation task. The **SemEval** column reports the statistics for the final set of records, i.e., gold records with double-agreement between annotators, and which we used to evaluate the systems. **Total** reports the statistics of all analyzed records, from which we chose our **SemEval** data. **Skipped** and **InProg** show the statistics for discarded records and records without a final decision, respectively. **Dev** shows the statistics for the development set.

Each of the rows reports a value of a component of the data or annotator interaction with the data. **Records** indicates the number of annotated verbs and their arguments. **Sentences** and **Tokens** indicate the size of the sub-corpus of the annotated records. **VF** is the number of distinct verb lemmas (273), mapped to the number of distinct frames that the **Frames-Type** row shows (149) (Figure 3 in Appendix A.1 plots their frequency distribution.) **FElements** reports the number of annotated FEs categorized under the number of FE types shown in the **FE-Type** row. **Sem-Arg** shows the number of annotated verb arguments with VerbNet-like semantic roles, classified into 32 of 41 possible semantic role categories. **Multi-word** lists the number of annotated MWUs

	SemEval	Total	Skipped	InProg	Dev
Records	4,620	5,637	301	716	594
Sentences	3,346	3,803	294	675	582
Tokens	90,460	102,067	8,329	19,151	15198
Verb-Forms	273	373	93	210	35
Frame-Type	149	234	75	185	37
#FEs	9,510	11,269	373	1,386	1,128
FE-Type	198	270	64	197	62
Sem-Arg	9,466	11,215	370	1,379	1,079
Multi-word	2,366	2,773	61	346	368
Confidence	3.30	3.2	2.41	2.5	3.34
Time	539h	742h	25h	177h	19h
Total-Move	68,784	83,753	1,903	13,066	4,406

Table 1: Annotation and Data Statistical Summary

Confidence reports the average of annotator-assigned confidence scores for annotations per record. Although interpreting this measure demands more work, the averages appear to be as expected. Specifically, **SemEval** is higher in value than both **InProg** and **Skipped**, facts that we associate with double agreement and the choice reviewing process. Still, many records with high confidence scores remained as **InProg** given the lack of double agreement. Table 5 (Appendix A.1) lists the top 10 frames annotated with their respective highest and lowest confidence ratings averaged by their frequency in **SemEval**.

The last two rows of Table 1 are meta-data on the annotation process. **Time** reports the total time annotators spent in active annotation, engaged in the steps described above (742 hours), excluding the reviewing process (Section 4.3.1) and including the time to annotate MWUs. **Total-Move** is the total number of logical moves for frame annotation between annotators and the annotation system, i.e., logged changes in the process of frame and core FE annotation. This number excludes annotation of verb subcategorization with generic semantic roles.[6]

In **SemEval**, annotated frames had an average of 2.15 arguments, requiring a minimum of five logical moves to annotate (MWU-less sentences). However, on average, each **SemEval** record required 14.8 moves. This number is even higher for **InProg** (18.2); we believe that it indicates the complexity of the annotation task. Table 4 (Appendix A.1) further details annotator activity, with time spent and moves per annotation step. As expected, frame annotation of verbs (Step 2), was the most time consuming part of the task.

4.5 Development Dataset

Shared task participants received a development set consisting of 600 records from a total of 4,620 records, where Table 4 shows the statistics. The development set contained gold annotations for all three subtasks.

5 Evaluation Metrics

For all subtasks, as figure of merit, here we report the performance of participating systems with measures for evaluating text clustering techniques, including the classic measures of Purity (PU), inverse-Purity (IPU), and their harmonic mean (PIF) (Steinbach et al., 2000), as well as the harmonic mean for BCubed precision and recall (i.e.,

[6]With the exception of a few verbs, annotators rarely changed the annotation system's rule-based suggestions of VerbNet semantic roles.

BCP, BCR, and BCF, respectively) (Bagga and Baldwin, 1998).

To compute these measures for the pairing of reference-labeled data and unsupervised-labeled data (with each having an exact set of annotated items), we built a contingency table T with rows for gold labels and columns for unsupervised system labels. We filled the table with the number of intersecting items, as done in cross-tabulation of results in classification tasks to compute precision and recall. For Task A (Section 3), T tracks the unsupervised system labels and the gold reference labels assigned to verbs. For Task B.1, we labeled the rows and columns of T with tuples (l_v, l_a), where l_v labels the frame evoking verb and l_a labels the FE filler. For Task B.2, the rows and columns in T track the unsupervised system labels and the gold reference labels (generic semantic roles) assigned to arguments.

These performance measures reflect a notion of similarity between the distribution of unsupervised labels and that of the gold reference labels, given certain criteria. Specifically, they define the notions of consistency and completeness of automatically generated clusters based on the evaluation data. Each method measures consistency and completeness in its own way, and alone may lack sufficient information for a clear understanding and analysis of system performance (Amigó et al., 2009). But, as the *single metric for system ranking*, we used the BCF measure, given its satisfactory behavior in certain situations. Note that we modeled the task and its evaluation as hard clustering, where a record receives only one label, without overlap in any generated category of items.

5.1 Baselines

Similar to other clustering tasks, we use baselines of random, all-in-one-cluster (AIN1), and one-cluster-per-instance (1CPI). Additionally, we adapted the baseline of *the most frequent sense* in WSI for these tasks by introducing the one-cluster-per-head (1CPH) baseline in Task A, and one-cluster-per-syntactic-category (1CPG) for verb argument clustering in Task B.2.[7] For Task B.1, we built a baseline, 1CPGH for labeling verbs with their lemmas (as in 1CPH) and FEs with grammatical relation to their heads (as in 1CPG). We included two more labels lcmpx and rcmpx for frame fillers with no direct syntactic relation to the head verb, if occurring left of or right of the verb, respectively.

Both 1CPH and 1CPG (and their combination for Task B.1) are hard to beat because of the long-tailed distribution of the frequency of our test data. E.g., most verbs frequently instantiate one particular frame and rarely other ones. Similarly, a particular role (FE) frequently is filled by words that have a particular grammatical relation to its governing verb; e.g., most subjects of most verb forms receive the agent label in their subcategorization frame (or, an agent-like element in their Frame Semantics representations). Evidently the chosen labels for grammatical relations influences 1CPG and 1CPHG scores. Values reported later (specifically, Tables 6 and 2) could be improved by employing heuristics, e.g., relabeling enhanced dependencies using a few rules.

We also employed one unsupervised and a second supervised system baselines. For the unsupervised one, we trained the system with data from Kallmeyer et al. (2018). For the supervised one, we used OPEN-SESAME, a state-of-the-art supervised FrameNet tagger (Swayamdipta et al., 2018). After converting its output to the format of the present task, we evaluated it similar to other systems. Both systems were trained out-of-the-box with no additional tuning.

6 System Descriptions

We received submissions from nine teams (13 participants). Only three chose to submit system description papers. Arefyev et al. (2019) provided a solution for Task A and Task B.2, using both sets of these results to address Task B.1. Task A used language models and Hearst-like patterns to tune and obtain contextualized vector representations for the verbs in the test set. A hierarchical agglomerative clustering method followed, where hyperparameters were set with labeled and unlabeled records from the development and test sets. Task B.2 employed a logistic regression trained over the development set to identify only the most frequent labels. The classifier was based on features obtained from a language model and hand-crafted rules. Using logistic regression and training this algorithm with the development set remains an issue of concern, given the intended unsupervised scenario. While we objected to using the development set to train a supervised system for this

[7] We use syntactic dependencies of the Enhanced Universal Dependencies formalism (Schuster and Manning, 2016).

subtask, we still report its scores. The differences between its results and those of the other systems may be informative. Still, we considered Arefyev et al.'s results for Task B only complementarily, not to rank the systems.

Anwar et al. (2019) proposed a method that was similar to that of Arefyev et al. (2019). Arefyev et al. used contextualized word embeddings from the BERT language modeling tool Devlin et al. (2018), whereas Anwar et al. used pre-trained embeddings. They merged the outputs of Tasks A and B.2 for Task B.1. Task A used agglomerative clustering of vectors with concatenated verb representation vectors and vectors that represent usage context. Task B.2 employed hand crafted features, a method to encode syntactic information, and again an agglomerative clustering method.

Ribeiro et al. (2019) also reported results for all subtasks using similar techniques to those reported in the other two submitted papers. Ribeiro et al. (2019) used the bidirectional neural language model BERT, which Arefyev et al. (2019) also used. Task A employed contextualized word representations proposed in (Ustalov et al., 2018), and Biemann's clustering algorithm (Biemann, 2006). Compared to the two other systems, Ribeiro et al. (2019) exploited input structures, weighted them, and used them elegantly in its algorithm. With the same method but different hyper-parameters for B.2 along with combining results from Task A, Ribeiro et al. (2019) offered a solution to B.1.

7 Results and Data Analysis

Table 2 reports the BCF scores for system submissions along with a baseline for each task.[8] As the table shows, each system performs best only in one of the tasks. We report Arefyev et al.'s submission for Tasks B.1 and B.2 only to show the benefit of using a small amount of training data and a supervised method together with a clustering algorithm, provided that such training data is available. As readers know, finding the optimal (actual) number of clusters is an open research area. Participants knew the number of clusters: whereas Arefyev et al. and Anwar et al. used this information, Ribeiro et al. opted for a statistical method tuned with data that we provided.

The baseline systems, the unsupervised method of Kallmeyer et al. (2018) performed the worst

[8]The full list of baselines and performance measures appear in Table 6 of the Appendix.

System	BCF	BCF	BCF
Arefyev et al.	**70.70**	~~63.12~~	~~64.09~~
Anwar et al.	68.10	**49.49**	42.1
Ribeiro et al.	65.32	42.75	**45.65**
BASELINE	65.35	45.79	39.03
	Task A	B.1	B.2

Table 2: Summary of Results. The BASELINE for Task A is 1CPH, and for B.1 and B.2 is 1CPHG. Best results appear in bold face; discarded results are crossed out. Table 6 lists all other baselines.

of all systems regarding BCF. This result is not surprising since that work did not effectively handle MWUs in the test, where only the head of the MWU was kept. However, the output of Open-SESAME, and its low BCF was indeed surprising.

We fed Open-SESAME the sentences from the test set; it identified approximately 5k frames. However, the overlap with the test set was only 1,216 records (identification problem in Open-SESAME). These 1,216 records exhibit a mismatch between 536 of the arguments and their respective target verbs. We ignored the system's extra or incorrectly generated arguments, and replaced the missing items with those of the 1CPHG baseline records. We then used the resulting records for evaluation against the task's gold data as did the task's participants. As Table 3 shows, the unsupervised method outperforms the supervised system for all tasks by a wide margin, i.e.,the unsupervised label set can carry more information than does the supervised label set.

	BCP	BCR	BCF
Task A	84.52	44.67	58.45
Task B.1	81.04	31.6	45.47
Task B.2	34.26	36.56	35.37

Table 3: Open-SESAME Performance

We compared results for confidence measure that annotators assigned to records. First, we split the evaluation records according to their assigned confidence value into five subsets E_i, $1 \leq i \leq 5$, such that subset E_1 contained only records with confidence value 1, E_2 contained only record with confidence value 2, etc.. Then we evaluated system outputs on each subset E_i and logged that BCF. Later, we performed this evaluation cumulatively using subsets E_i's by adding records from all E_js to E_i where $i < j$. Interpreting the obtained values requires careful attention (e.g., changes in the prior probabilities of gold clusters and their

cardinality must be taken into account), overall, we observed a similar trend for all systems: as expected, namely a positive correlation between the confidence value and BCF. Thus, what human annotators usually found hard to annotate, automatic systems also found hard to cluster. (The reverse relation does not hold). Or, pessimistically, the level of noise in annotation increases as their associated confidence decreases. (Table 7 in Appendix A.2 details the results.)

Finally, we wanted to identify the frames that machines found difficult to cluster. To estimate difficulty we used the differences in BCF under the following conditions. We repeated the evaluation process $1 \leq i \leq n$ times (where n is the number of gold labels for a task) for each system. In each iteration i, we removed all data items of a gold category i. We measured and noted the resulting BCF in the given iteration; we deducted the score from the system performance over the entire gold set. To cancel frequency effects, we normalized the differences by the number of gold data instances. We removed all records annotated as COMMERCE_SELL from the evaluation set E to form E'. We computed the BCF of the systems over E' ($E' \subset E$), and measured $d = E_{\text{BCF}} - E'_{\text{BCF}}$. We interpreted a positive difference as an easy to cluster gold category i, and a negative difference as a hard to cluster gold category i.

The heat maps in Table 8 and Table 9 show a summary of the results for Task A and Task B.2, respectively. All systems performed similarly for approximately 30% of the gold classes. Comparing differences across systems and the baselines of 1CPH and 1CPG reveals (possibly) interesting information. Thus, for example, in Task A, most systems found COMMERCE_SELL hard and COMMERCE_BUY easy to cluster. Interestingly, a set of six verbs evokes each frame: *buy, purchase, buy_back, buy_up, buy_out, buy_into* for COMMERCE_BUY; and *sell, retail, auction, place, deal, resell* for COMMERCE_SELL. From these two sets of verbs, three are polysemous: *buy* in the former, and *place* and *deal* in the latter. Does the morphology of the verbs (e.g., *buy-back*, *resell*) make one easy to cluster? Alternatively, are other factors at play, such as the number of verb instances? How these factors might influence the proposed naive BCF-difference model is an open question.

8 Concluding Remarks

We have presented the SemEval 2019 task on unsupervised lexical frame induction. We described the task in detail, provided a summary of methods that participants developed, and compared the results. Although much room for improvement of the task remains, we consider it a step forward. It employed a well-motivated typology of lexical frames to distinguish lexical frame induction tasks. The evaluation data derived from annotations of a well-known resource, namely a portion of WSJ sentences, perhaps the most annotated corpus of English. These features provide opportunities for future investigation, in particular in studies related to reciprocal relations between syntactic and lexical semantic frame structures.

One reason to promote using unsupervised methods is their inherent flexibility to embrace unknown data. These methods have a high margin of tolerance for *noise*, and perform better than supervised method with insufficient training data. For unsupervised data, obtaining or generating training data is easier than doing so with supervised methods because they simply do not require annotation. For example, all participant systems could collect similar unlabeled training data from only syntactically annotated corpora to generate more unlabeled records. Ultimately, such methods can achieve respectable performance, and produce clusters which are both more informative than the unlabeled input and supervised categories (under certain situations). As shown, unsupervised methods can even outperform a state-of-the-art Frame Semantics parser by a wide margin (Section 7), while a very large gap remains for improvements in future work.

Acknowledgements

This research was funded by DFG - SFB991. We thank Timm Lichte, Rainer Oswald, Curt Anderson, and Kurt Erbach. We also thank the LDC for its generous support, and the NVIDIA Corporation for the Titan Xp GPU used in this work.

References

Eneko Agirre and Aitor Soroa. 2007. Semeval-2007 task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of the 4th International Workshop on Semantic Evaluation*, SemEval '07, pages 7–12, Stroudsburg, PA, USA. Association for Computational Linguistics.

Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Inf. Retr.*, 12(4):461–486.

Saba Anwar, Dmitry Ustalov, Nikolay Arefyev, Simone Paolo Ponzetto, Chris Biemann, and Alexander Panchenko. 2019. Hm^2 at semeval 2019 task 2: Unsupervised frame induction using contextualized and uncontextualized word embeddings. In *Proceedings of The 13th International Workshop on Semantic Evaluation*.

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*.

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the 17th International Conference on Computational Linguistics - Volume 1*, COLING '98, pages 79–85, Stroudsburg, PA, USA. Association for Computational Linguistics.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186. 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. Association for Computational Linguistics.

Claire Bonial, Kevin Stowe, and Martha Palmer. 2013. Renewing and revising semlink. In *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and linking lexicons, terminologies and other language data*, pages 9 – 17, Pisa, Italy. Association for Computational Linguistics.

Silvie Cinková, Eva Fučíková, Jana Šindlerová, and Jan Hajič. 2014. EngVallex - English valency lexicon. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

C. 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.

Charles J. Fillmore. 1968. The case for case. In *Universals in Linguistic Theory*, pages 1–88. Holt Rinehart and Winston, New York.

Thomas Gamerschlag, Doris Gerland, Rainer Osswald, and Wiebke Petersen, editors. 2014. *General Introduction*. Springer International Publishing, Cham.

Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. 2007. Semeval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of the Fourth International Workshop on Semantic Evaluation (SemEval-2007)*, pages 13–18. Association for Computational Linguistics.

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 on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. 2012. Announcing prague czech-english dependency treebank 2.0. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*. European Language Resources Association (ELRA).

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*, volume 1, pages 471–482.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, NAACL-Short '06, pages 57–60, Stroudsburg, PA, USA. Association for Computational Linguistics.

David Jurgens and Ioannis Klapaftis. 2013. Semeval-2013 task 13: Word sense induction for graded and non-graded senses. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pages 290–299.

Laura Kallmeyer, Behrang QasemiZadeh, and Jackie Chi Kit Cheung. 2018. Coarse lexical frame acquisition at the syntax–semantics interface using a latent-variable pcfg model. In *Proceedings of the Seventh*

Joint Conference on Lexical and Computational Semantics, pages 130–141, New Orleans, Louisiana. Association for Computational Linguistics.

Karin Kipper, Hoa Trang Dang, and Martha Palmer. 2000. Class-based construction of a verb lexicon. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 691–696. AAAI Press.

Alessandro Lenci and Giulia Benotto. 2012. Identifying hypernyms in distributional semantic spaces. In *SemEval 2012*, pages 75–79. Association for Computational Linguistics.

Suresh Manandhar, Ioannis Klapaftis, Dmitriy Dligach, and Sameer Pradhan. 2010. Semeval-2010 task 14: Word sense induction & disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 63–68, Uppsala, Sweden. Association for Computational Linguistics.

Jiří Materna. 2012. Lda-frames: An unsupervised approach togenerating semantic frames. In *Computational Linguistics and Intelligent Text Processing*, pages 376–387, Berlin, Heidelberg. Springer Berlin Heidelberg.

Diana McCarthy and Roberto Navigli. 2007. Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the Fourth International Workshop on Semantic Evaluation (SemEval-2007)*, pages 48–53. Association for Computational Linguistics.

George A. Miller. 1995. WordNet: A lexical database for English. *Commun. ACM*, 38(11):39–41.

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.

Roberto Navigli and Daniele Vannella. 2013. Semeval-2013 task 11: Word sense induction and disambiguation within an end-user application. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 193–201, Atlanta, Georgia, USA. Association for Computational Linguistics.

Stephan Oepen, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Silvie Cinkova, Dan Flickinger, Jan Hajic, Angelina Ivanova, and Zdenka Uresova. 2016. Towards comparability of linguistic graph banks for semantic parsing. In *LREC 2016*, Paris, France. ELRA.

Martha Palmer, Claire Bonial, and Jena Hwang. 2017. Verbnet: Verbnet: Capturing english verb behavior, meaning, and usage. In *The Oxford Handbook of Cognitive Science*. Oxford Press.

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

Marco Pennacchiotti, Diego De Cao, Roberto Basili, Danilo Croce, and Michael Roth. 2008. Automatic induction of framenet lexical units. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 457–465, Stroudsburg, PA, USA. Association for Computational Linguistics.

Behrang Q. Zadeh and Miriam R. L. Petruck. 2019. Guidelines for the semantic frame annotation system. corpus annotation guidelines TR.9.2018, SFB991 - ICSI.

Drew Reisinger, Rachel Rudinger, Francis Ferraro, Craig Harman, Kyle Rawlins, and Benjamin Van Durme. 2015. Semantic proto-roles. *Transactions of the Association for Computational Linguistics*, 3:475–488.

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*.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, Collin F. Baker, and Jan Scheffczyk. 2016. *FrameNet II: Extended Theory and Practice*. ICSI, Berkeley.

Sebastian Schuster and Christopher D. Manning. 2016. Enhanced English Universal Dependencies: An improved representation for natural language understanding tasks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

M. Steinbach, G. Karypis, and V. Kumar. 2000. A comparison of document clustering techniques. In *KDD Workshop on Text Mining*.

Swabha Swayamdipta, Sam Thomson, Kenton Lee, Luke Zettlemoyer, Chris Dyer, and Noah A. Smith. 2018. Syntactic scaffolds for semantic structures. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3772–3782, Brussels, Belgium. Association for Computational Linguistics.

Dmitry Ustalov, Alexander Panchenko, Andrey Kutuzov, Chris Biemann, and Simone Paolo Ponzetto. 2018. Unsupervised semantic frame induction using triclustering. In *ACL*, pages 55–62, Melbourne, Australia. ACL.

A Appendices

A.1 Appendix I: Annotation Process

A.1.1 Time and Moves per Annotation Step

Table 4 shows the amount of effort to develop the SemEval dataset in terms of time and moves that the annotation system recorded. (See Sections 4.3, 4.4).

Annotator Activity	Time	Moves
Reading and Comprehension	78	4,795
Choosing a Frame	177	9,737
Annotating Arguments	81	19,510
Rating, Revising, Commenting	115	25,793
Multi-word Unit Annotation	89	8,949
Total	539	68,784

Table 4: Total hours and number of moves for each annotation step for the 4,620 record dataset.

A.1.2 Plot of frequency of annotated frames

Figure 3 plots the frequency distribution of the annotated frames in the gold data (**SemEval**).

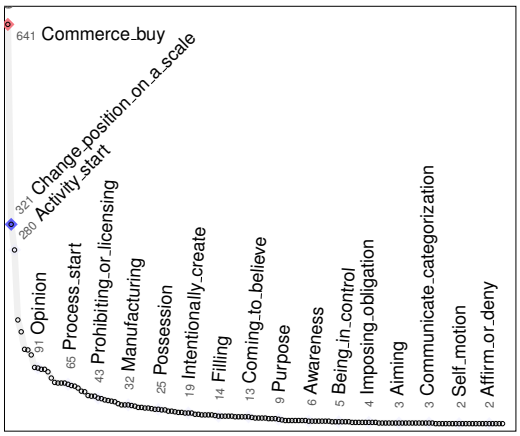


Figure 3: Frequency Distribution of Annotated Frames

A.1.3 Some Frames and their Averaged Confidence

Table 5 lists FN frames annotated with the highest and lowest confidence. Table 4 details hours spent to derive the evaluation data set. Section 4.3 discusses both tables. The full list of annotations in human readable form is available to browse and comment on at http://corpora.phil.hhu.de/fi/frames.html.

A.2 Appendix II: Statistical Summary of Evaluation and System Submissions

A.2.1 Unabridged Results Table

Table 6 extends Table 2. Section 5 defines the abbreviations. A horizontal line separates participating systems and the baselines.

A.2.2 Confidence Measures and BCF Performance

Table 7 shows system BCF scores for confidence. The table shows changes in the BCF of systems when altering the evaluation set based on the assigned confidence for an annotated record. (See Section 7 for an explanation).

Frame Type	#VF	#Rec	Conf
DECIDING	1	13	4.31
AGREE_OR_REFUSE_TO_ACT	1	15	4.13
TAKE_PLACE_OF	1	11	4
BEING_EMPLOYED	1	6	4
STATEMENT	8	149	3.97
TAKING_SIDES	3	16	3.88
ACTIVITY_STOP	4	16	3.88
COMMERCE_SELL	6	168	3.82
BRINGING	1	5	3.8
GIVE_IMPRESSION	4	39	3.79

(a) Frames with Highest Average Confidence

Frame Type	#VF	#Rec	Conf
BEING_IN_CONTROL	2	5	1.6
COMING_TO_BE	2	5	1.8
OPERATING_A_SYSTEM	2	10	1.8
AWARENESS	1	6	1.83
REMOVING	3	8	1.88
INTENTIONALLY_CREATE	6	19	1.95
CERTAINTY	1	68	2.03
OPINION	2	91	2.1
THWARTING	2	22	2.32
FIRST_RANK	1	21	2.38

(b) Frames with Lowest Average Confidence

Table 5: Frame types with the highest (5a) and the lowest (5b) confidence (**Conf**) by number of records (**#Rec**) with double annotator agreement. **#VF** reports the number of distinct verb forms that evoke a frame.

System	#C	PU	IPU	PIF	BCP	BCR	BCF
Arefyev et al.	272	**78.68**	77.62	**78.15**	**70.86**	70.54	**70.7**
Anwar et al.	150	72.4	**81.49**	76.68	62.17	**75.27**	68.1
Ribeiro et al.	222	72.84	77.84	75.25	61.25	69.96	65.32
Kallmeyer et al.	218	73.77	72.86	73.31	64.62	65.48	65.05
1CPI	4620	100	3.23	6.25	100	3.23	6.25
AIN1	1	13.87	100	24.37	3.78	100	7.28
1CPH	273	82.16	66.95	73.78	75.98	57.33	65.35
RANDOM	149	15.11	5.78	8.36	6.76	3.85	4.9

Task A

System	#C	PU	IPU	PIF	BCP	BCR	BCF
Arefyev et al.	~~776~~	~~72.47~~	~~72.16~~	~~72.31~~	~~62.73~~	~~63.51~~	~~63.12~~
Anwar et al.	338	55.74	**67.79**	**61.18**	43.22	**57.9**	**49.49**
Ribeiro et al.	518	52.29	57.56	54.8	39.43	46.69	42.75
Kallmeyer et al.	1023	**72.24**	49.12	58.48	**62.71**	37.51	46.94
1CPI	9510	100	4.58	8.77	100	4.58	8.77
AIN1	1	6.55	100	12.3	1.56	100	3.08
1CPHG	1203	78.46	45.99	57.99	71.11	33.77	45.79
RANDOM	436	11.34	6.04	7.88	6.03	4.81	5.35

Task B.1

System	#C	PU	IPU	PIF	BCP	BCR	BCF
Arefyev et al.	~~14~~	~~73.94~~	~~81.4~~	~~77.49~~	~~56.25~~	~~74.46~~	~~64.09~~
Anwar et al.	2	50.43	**80.47**	**62.00**	29.58	**73.00**	42.1
Ribeiro et al.	7	58.25	71.4	64.16	36.88	59.91	**45.65**
Kallmeyer et al.	37	**61.44**	51.53	56.05	**40.89**	37.33	39.03
1CPG	37	61.44	51.53	56.05	40.89	37.33	39.03
1CPI	9466	100	0.34	0.67	100	0.34	0.67
AIN1	1	34.34	100	51.13	21.66	100	35.6
RANDOM	32	34.65	4.75	8.36	21.89	3.45	5.96

Task B.2

Table 6: Complete System Results and Baselines

Cnf	#I	Arefyev	Anwar	Ribeiro
1	4620	70.7	68.10	65.32
2	4334	71.87	69.28	66.57
3	3657	74.64	72.22	70.17
4	2542	76.46	73.82	73.43
5	84	86.14	84.65	85.13

Task A

Cnf	#I	Arefyev	Anwar	Ribeiro
1	9,510	63.12	49.52	42.75
2	9017	64.20	50.44	43.61
3	7,606	67.18	53.40	46.42
4	5,356	68.70	55.99	49.20
5	169	85.16	81.85	65.60

Task B.1

Cnf	#I	Arefyev	Anwar	Ribeiro
1	9,466	64.09	42.12	45.65
2	8,911	64.98	42.32	46.27
3	7,528	66.47	42.67	47.52
4	5,292	65.71	40.67	46.95
5	167	77.19	55.18	56.58

Task B.2

Cumulative

Cnf	#I	Arefyev	Anwar	Ribeiro
1	286	73.79	70.57	67.70
2	677	66.45	63.80	60.46
3	1,115	76.71	75.98	70.01
4	2,458	76.65	74.05	73.45
5	84	86.14	84.65	85.13

Task A

Cnf	#I	Arefyev	Anwar	Ribeiro
1	493	68.57	55.37	51.84
2	1,411	59.86	49.08	42.16
3	2,250	70.67	57.97	47.60
4	5,187	68.70	56.01	49.24
5	169	85.16	81.85	65.60

Task B.1

Cnf	#I	Arefyev	Anwar	Ribeiro
1	553	52.69	39.82	38.21
2	1,385	58.36	40.99	41.55
3	2,236	69.01	48.07	49.4
4	5,125	65.44	40.37	46.72
5	167	77.19	55.18	56.58

Task B.2

Stratified

Table 7: Changes in BCF score of systems relative to changes in evaluation records based on assigned confidence measure.

A.3 Examining Clusters by Removing One Gold Cluster at a Time

#Rmvd	Frame	Arefyev	Anwar	Ribeiro	Kallmeyer	1CPH
168	Commerce_sell					
6	Awareness					
89	Assessing					
54	Seeking_to_achieve					
30	Activity_ongoing					
19	Intention._creat.					
2	Storing					
3	Cause_..._progress					
27	Process_continue					
14	Filling					
121	Causation					
59	Hiring					
89	Choosing					
6	Being_employed					
7	Criminal_investig.					
65	Process_start					
4	Notifi..._charges					
28	Assistance					
21	First_rank					
20	Giving					
2	Sentencing					
16	Activity_stop					
641	Commerce_buy					
42	Have_associated					
13	Request					
149	Statement					
67	Avoiding					
2	Performer..._roles					
2	Attack					
92	Cause_change...					
9	Purpose					
14	Collaboration					
13	Coming_to_believe					

Table 8: Task A – Part of a heat map from results (Section 7), with cases that exhibit a range of difference values. Red denotes a positive and blue a negative difference; white means no change (zero difference). Differences (normalized by cluster size) are in domain 0.01 to −0.01.

#Rmvd	Frame	Arefyev	Anwar	Ribeiro	Kallmeyer	1CPH
2	Attending					
38	Getting					
16	Activity_stop					
63	Becoming_a_member					
641	Commerce_buy					
6	Retaining					
7	Criminal_investigatic					
67	Avoiding					
46	Scrutiny					
280	Activity_start					
5	Bringing					
2	Change_event_time					
89	Choosing					
8	Removing					
13	Coming_to_believe					
14	Inspecting					
3	Cause_to_end					
3	Communicate_categ					
13	Deciding					
2	Attack					
2	Creating					
19	Intentionally_act					
34	Cause_to_amalgamat					
19	Intentionally_create					
5	Usefulness					
16	Taking_sides					
4	Notification_of_charg					
3	Aiming					
121	Causation					
3	Process_end					
2	Transfer					
5	Coming_to_be					
10	Cotheme					
7	Hearsay					
16	Transition_to_state					
92	Cause_change_of_pos					
7	Inclusion					
7	Simultaneity					
4	Imposing_obligation					
3	Cause_change					
3	Distributed_position					
39	Give_impression					
4	Supporting					

Table 9: Heat map that visualizes Task B.2 data