Unsupervised Lexical Semantic Frame Induction: A SemEval 2019 Task Proposal

Behrang QasemiZadeh[1], Miriam R. L. Petruck[2], Laura Kallmeyer[1],
Marie Candito[3], Alfredo Maldonado[4], Rainer Osswald[1], and Timm Lichte[1]

[1] University of Düsseldorf

[2] International Computer Science Institute

[3] Paris Diderot University

[4] Trinity College Dublin

1 Motivation

Frame Semantics (Fillmore, 1976) and other theories (Gamerschlag et al., 2014) that adopt typed feature structures for the representation of knowledge and linguistic structures have developed in parallel over several decades in formal linguistics studies related to the syntax–semantics interface, as well as in empirical corpus-driven applications in natural language processing (NLP). Building repositories of lexical semantic frames is a central topic in these efforts, regardless of their perspective. In formal studies, lexical semantic frame knowledge bases are built to instantiate foundational theories with tangible examples, e.g., to serve as supporting evidence for the theory. On a practical level, frame semantic repositories play a pivotal role in natural language understanding and semantic parsing (both as a source for inspiring a representation format and for training data-driven machine learning methods) to accomplish tasks such as information extraction, question answering, text summarization, machine translation, and so on.

However, the manual development of lexical semantic frames databases (as well as corpus-derived annotations to support those frames) is a resource-intensive task. The most well-known publicly available Frame Semantics lexical resource is FrameNet (Ruppenhofer et al., 2016), which covers only a fraction of natural language events and concepts in a relatively small number of contextual semantic domains. While NLP research has integrated FrameNet into semantic parsing technologies (e.g. Das et al. (2014); Kshirsagar et al. (2015); Hartmann et al. (2017)), current parsing methods are not yet sufficiently effective for enriching frame repositories such as FrameNet with new frame templates, i.e., to port them to new semantic domains, and to extend them to languages other than English. In general, the same holds for supervised machine learning for the identification of frames. One way to rectify this situation is the use of unsupervised machine learning methods for identifying new semantic frame templates and populating them. Even if these unsupervised approaches are not ideal to create full-fledged frame semantic databases, employing them as an assistive lexicographical tool might well reduce the resource intensivity of the effort.[1] Among the studies of unsupervised frame induction, most systems, e.g., Pennacchiotti et al. (2008) and Green et al. (2004), address the problem of domain coverage by employing other available lexical semantic resources such as WordNet (Miller, 1995), a technique that itself leads to other setbacks. Most importantly, these methods do not adapt easily across languages since for the most part these auxiliary resources are often not available in other languages.

[1] Of course, these unsupervised methods may have applications other than lexicography and lexical semantic studies, e.g., cold start knowledge base construction (Roth et al., 2015) and populating ontologies from text (Buitelaar, 2005).

The ambitious goal of this task is to induce frame structures automatically using **un**supervised methods from corpora that are only annotated with grammatical relationships in the universal dependencies (UD) formalism.[2] Specifically, as the first step, the goal is to identify the event frames that verbs evoke, and then to cluster the verbs and their syntactic argument(s) into high level frame clusters and semantic argument groups. For example, given sentences such as (1)–**??**, below, the goal is to group the verbs into frames such as those in Table 1[3], and then further group the arguments of verbs into frame specific slot fillers clusters, such as `items`[4], `attributes`[5], `difference`[6], and `final_value`[7] for (1) and (2) as listed in Table 2.

(1) Share prices *plummeted* to new lows in value.

(2) Total tax revenue *increased* by 5.1.

(3) The gallery was *packed* with Judge Hastings's supporters.

For

`Change position on a scale`: plummeted (1), increased (2)
`Filling`: packed (3)
`Placing`: packed **??**

Table 1: Grouping verbs into frame clusters.

`items`: Share prices (1), Total tax revenue (2)
`attributes`: value (1)
`difference`: 5.1 (2)
`final_value`: new lows (1)

Table 2: Slot filler clusters for the `Change position on a scale` frame from (1) and (2).

Note that unlike efforts that aim to build a semantic graph from natural language sentences (Dohare and Karnick, 2017; Damonte and Cohen, 2017; May and Priyadarshi, 2017), here, the goal is to induce semantic frames for individual verbs in sentences. Thus, for example, in (4) the task will not deal with the discourse-related (cause-effect) phenomena between *plummet* and *being hosted*.

(4) Top Gear ratings *plummet* in US despite being hosted by American star Matt LeBlanc.

To the best of our knowledge, the proposed unsupervised task has not yet been the subject of an evaluation effort such as SemEval (or its predecessor Senseval). The task differs from the previous shared tasks on word sense induction and unsupervised role labelling in several ways. Like other word-sense induction (WSI) shared tasks, (e.g., Agirre and Soroa (2007); Manandhar et al. (2010); Jurgens and Klapaftis (2013); Navigli and Vannella (2013)), grouping verbs into frame categories requires identifying polysemous words. However, the proposed task goes beyond WSI since it demands identifying lexical semantic

[2] Note that assuming the availability of UD corpora is well within reason, since the UD project already covers more than 62 languages: `http://universaldependencies.org/`.

[3] We adapt frame definitions from FrameNet, see `https://framenet.icsi.berkeley.edu/fndrupal/frameIndex`.

[4] The entity that has a position on the scale.

[5] The scalar property that the item possesses.

[6] The distance by which an Item changes its position on the scale.

[7] The position on the scale where the `item` ends up.

relationships such as synonymy, hyponymy (troponymy), meronymy, and antonymy, too (see the examples in Table 1). While previous shared tasks have addressed the identification of lexical semantic relations (e.g., Camacho-Collados et al. (2017); Jurgens et al. (2014); Girju et al. (2007); Hendrickx et al. (2009); Jurgens et al. (2012)), these were mostly supervised, lexical relations were treated in isolation, and the proposed evaluation was usually limited to identifying relations between nominals. In contrast, the current proposed task deals with verbs and the lexical semantic relations among them in an unsupervised setting for the goal of coarse lexical frame induction. Lastly, while a great deal of research has focused on unsupervised role labelling/clustering (e.g., Baisa et al. (2015); Lang and Lapata (2011); Surdeanu et al. (2008); Swier and Stevenson (2004)), in most of them, information about word senses is assumed to be known and given as input. In contrast, here the aim is to find clusters of verb arguments at the same time as inducing verb groups for each frame.[8] As indicated, none of the above-mentioned evaluation efforts were tailored to meet the needs of frame identification, particularly, for lexicographic resource development.

The data set for evaluating the current task consists of sentences with FrameNet annotation with the additional role-like layer. Specifically, we have annotated a new subset of sentences with FrameNet frames, for which preliminary annotation exists on a substantial number of sentences from the WSJ corpus; these data require a second pass to check the annotation. Compared to the available FrameNet annotated corpora, i.e. FrameNet's "full-text" annotation, we changed some of the argument and frame categories for two reasons: (a) to fit a clustering task, i.e., to ensure that all frames reflect concepts at a similar level of abstraction, and (b) as mentioned, to provide role-like groupings of slot fillers to make them more useful for applications such as question-answering. These provisions ensure the quality of the evaluation dataset; although a large number of frame annotated sentences are publicly available, the current evaluation data have not been accessible prior to this evaluation.

1.1 Proposed Task and Expected Impact

This proposal constitutes the first shared task on the problem of unsupervised coarse semantic frame induction. Grouping verbs into frames requires distinguishing between/among different senses of those verbs, and identifying the lexical semantic relations, e.g. synonymy and troponymy, that (may) hold between them. Also, syntactic arguments of verbs must be clustered as frame elements, a task that itself requires distinguishing between/among different senses of verbs and the nominals that fill their slots. Among the important advantages of this proposal are:

1. Employing a well-motivated typology of event frames to study a problem for which no previous evaluation task exists sets the stage for new lines of research. The envisaged annotation layers provide both fine-grained and coarse-grained categorization for the semantic grouping of verbs and their argument structure.

2. The data derive from annotations of a well-known resource, namely a portion of WSJ sentences. These sentences were annotated for other types of analyses; consequently, they provide opportunities for future investigation of reciprocal relationships between frame structures, as well as other types of linguistic analysis (e.g., sense disambiguation, syntactic analysis, discourse analysis, etc.).

3. The task demands addressing and evaluating a number of well-known problems in semantic evaluation efforts for lexical acquisition at the syntax-semantics interface, e.g.,

[8] To the best of our knowledge, Modi et al. (2012) is the only attempt in this direction.

sense disambiguation, lexical semantic relationship identification, argument structure mining, etc. These data (and potentially the methods) can be used for evaluating with respect to any of the aforementioned sub-problems, and with respect to the proposed problem of semantic frame acquisition. Doing so will facilitate understanding the interaction between/among solutions for the sub-problems in a complete and comprehensive setting.

We expect the task to attract researchers with interests in a broad range of topics, such as computational linguistics and NLP, knowledge acquisition, and the semantic web, to name but a few. Similarly, we anticipate interest from researchers investigating a broad range of machine learning methods, including traditional clustering methods and Bayesian networks as well as more recent deep neural network techniques and auto-encoders. Specifically, the task poses a challenge to the deep learning research community, since these methods have rarely been used for unsupervised induction of complex lexical-semantic structures. We anticipate a major impact on the field in comparing these methods with more traditional ones such as Bayesian networks. The current task also seeks to foster exchange between these communities. Most importantly, the long-term goal of the task is to build a gold-standard benchmark data set and baselines for a task that so far has not received much attention.

2 Task Description

For a number of syntactically parsed sentences and the given verbs[9] with the specified syntactic arguments, the participants will:

- Group the verbs into different clusters, each resembling the event frame that the verbs evoke, according to the gold standard data.
- Cluster the syntactic arguments such that each cluster resembles the slot fillers that they evoke. Ideally, syntactic arguments that fill the same role/slot will be grouped together.

Accordingly, participants will perform the following two sub-tasks:

- **Sub-Task 1:** Cluster verbs into frame groupings.
- **Sub-Task 2:** Groupings of both verbs and their syntactic arguments into the desired latent semantic clusters.

The task is unsupervised in the sense that except for syntactic annotations, no explicit semantic annotations (e.g., named entities) are allowed. However, participants are encouraged to use unsupervised computational learning methods, e.g., word embeddings, brown clusters, etc., to elicit such information from the training corpus and the auxiliary large raw-text web corpora that the organizers will provide. Participants may use syntactic/discourse parses other than the provided UD parses[10].

2.1 Evaluation Metrics

As figure-of-merit, we report the performance of participating systems using a range of measures often used for the evaluation of clustering techniques. These include the classic measures of:

[9] Other than copulas and semi-copulas.

[10] With the condition that their use be acknowledged explicitly.

- Purity (pu), inverse-Purity (ipu) and their harmonic mean (fpu) proposed in (Steinbach et al., 2000);
- the harmonic mean of BCubed's precision and recall (denoted by bcp, bcr, and bcf respectively) (Bagga and Baldwin, 1998);
- the adjusted Rand index (ari) metric (Rand, 1971), and
- an edit distance based measure (dst) proposed by Pantel and Lin (2002).

These measures reflect a notion of similarity between the distribution of instances in the obtained clusters and the gold/evaluation data based on certain criteria, roughly, by defining the notions of 'consistency/homogeneity' and 'completeness' of automatically generated clusters w.r.t. to the gold data. Each method has its own way of measuring 'consistency and completeness' and alone may lack sufficient information for a fair understanding and analysis of the systems' performances, as described in Amigó et al. (2009). However, as the single metric for the final ranking of the systems we will use the bcf measure.

For sub-tasks 1 and 2, we report these measures for grouping verbs into their frame clusters (both sub-tasks 1 and 2) and for clustering syntactic argument instances to their corresponding semantic slot/role groupings (task 2 only). Note that we design the task as a hard clustering in which the induced clusters by the systems are *not* overlapping.

2.2 Baselines

Apart from the random, all-in-one-cluster (AllIn1) and one-cluster-per-instance (1CpI) baselines, we report the performances obtained from our pilot systems (one based on a hierarchical Bayesian network). Moreover, we adapt the baseline of the most-frequent sense in WSI for our tasks by introducing the one-cluster-per-lexical-head (1CpH) and one-cluster-per-syntactic-category (1CpG) baselines for verb and argument clustering, respectively. Similar to WSI tasks, in our pilot tests, we found both 1CpH and 1CpG particularly challenging to beat due to the long-tailed distribution of lexical items in frame and role groupings, i.e., most verbs frequently instantiate one particular frame and rarely other ones (very similar to the distribution of words and their senses). Similarly, we observe that a particular role/slot frequently is filled by words that are in particular grammatical relation to the verb that evokes the frames, i.e., most subjects of verbs are the 'agent' slot/filler of the frame that the verbs evoke (in other words, the long-tailed distribution of subcategorization frames associated with a particular frame).

3 Dataset and Annotation

We derive and build the evaluation dataset from 1,000 sentences[11] chosen randomly from the PTB's WSJ sections.[12] As part of the training/evaluation dataset, we provide automatic parses of these sentences in the UD enhanced format (Schuster and Manning, 2016).[13]

[11] This number may increase slightly by the final deadline set for the evaluation campaign.

[12] https://catalog.ldc.upenn.edu/LDC2016T10.

[13] Despite their availability, we will not provide the gold-standard syntactic annotations as part of the training data to consider the effect of noise in the overall output. However, the availability of the gold-standard syntactic parses allows for future error analysis. In addition, these sentences have been annotated for semantic roles (Oepen et al., 2016) previously, where the verbs are also disambiguated and grouped according to the schema proposed in (Cinková et al., 2014). These annotations can also be helpful for future studies where sense groupings of lexical items can be compared with the FrameNet grouping prepared for this proposed shared-task.

#s4									
1	Papers	paper	_	NNS	_	3	nsubjpass	_	_
2	are	be	_	VBP	_	3	auxpass	_	_
3	packed	pack	_	VBN	_	0	ROOT	_	_
4	on	on	_	IN	_	6	case	_	_
5	the	the	_	DT	_	6	det	_	_
6	desk	desk	_	NN	_	3	nmod:on	_	_
7	.	.	_	.	_	-1	null	_	_

Table 3: The UD parse for example **??**; the sentence is assigned to the id #s4.

The major portion of the annotation guidelines for this task are based on those used for developing FrameNet, i.e. frame/slot definitions are borrowed from FrameNet. However, for argument annotation, a) we use a set of head projection rules, i.e., phrase-level arguments are annotated by marking words that are the head of these phrases; and b) we limit argument annotations to core elements only.

Frame annotations for the chosen sentences are presented in a style similar to that of Ontonote and Propbank. That is, one or more annotation records will accompany each sentence. Each annotation record consists of: (i) the id of the sentence from which the annotation comes; (ii) the position of the verb for which the sentence is annotated; (iii) the type of the frame assigned to the verb; and (iv) the arguments of the frame, i.e., one or more triples consisting of (a) a lexical filler; (b) its position in the sentence, and (c) its assigned slot filler type. That is, each annotation record has the following strucutre:

#sent-id position verb-lemma.**ftype** (word, position, **slot-type**)$^{+}$

For instance, for the sentence in **??** above (with #s4 as its id), the dataset will consist of its UD parse in the CoNLL-U format (with Penn-style part-of-speech tags), as seen in Table 3. Also, the dataset includes the following annotation record for the verb *packed*:

#s4 3 pack.**Placing** (paper, 1, **Theme**) (desk, 3, **Goal**)

During the evaluation period, a small trial dataset consisting of gold-standard annotation will be published. For testing, participants will receive the annotated structures *without* the labels assigned to frames or slot/role fillers, e.g. (for **??**):

#s4 3 pack.**UnKnown** (paper, 1, **UnKnown**) (desk, 3, **UnKnown**)

We expect participating systems to replace UnKnown labels with the unique cluster labels that they generate.

Annotation Process and Quality Assurance Checks: Each sentence in the dataset will be annotated by at least two annotators.[14] Since our goal is to build a gold-standard benchmark, we cross-check the manual annotations that different annotators have provided. The task organizers and the annotators discuss inconsistent cases until they reach an agreement.[15] Hence, the aim is to achieve a 100% inter-annotator agreement. However, we will report and discuss lessons learned during the annotation process in our task paper (e.g., changes in inter-annotator agreement, methods for resolving inconsistencies, etc.).

[14] Currently two trained linguists are working on the data.

[15] In the worst case, the problematic instance will be removed (a similar procedure previously employed in developing word-sense annotated datasets).

FT	FI	V	AT	AI
27	5,324	169	46	7261

Table 4: Current statistics of gold-standard data development: FT, FI, V, AT, and AI denote the number of frame types, instances, distinct verb heads, argument types, and argument instances that are annotated so far. Note that these statistics are subject to change.

Frame	#T	#V	{Examples of verbs occurrences}
Change_position_on_sacle	1259	17	{*fall*:356, rise:271, drop:135, decline:119, ...}
Activity_start	290	2	{*begin*:182, start:108}
Process_start	188	2	{*begin*:143, start:45}
Placing	121	21	{*place*:62, pack:3, wrap:1...}
Scrutiny	77	9	{*investigate*:28, examine:16, search:3 ...}
Filling	35	12	{*fill*:14, pack:6, cover:3, wrap:2 ...}
Adorning	26	10	{*fill*:8, cover:4, adorn:2, dot:2, encircle:2, ...}
Seeking	11	3	{*search*:9, hunt:1, probe:1}

Table 5: Examples of frames annotated to date and verbs that evoke them. #T and #V denote the total number of instances for the frame and the number of distinct verb lemmas that evoke them, respectively. Examples of verb lemmas and their occurrence frequency are listed in the last column.

Current state of annotation process: The 1,000-sentence dataset is in active development. So far, all the verbs in these sentences have been assigned to a FrameNet frame by one annotator, with the second annotator currently working on this task. Additionally, the core arguments of these frames are partially annotated by the first annotator. Table 4 reports statistics for annotations completed to date. Table 5 provides examples of annotated frame instances and verbs that evoke them. The complete list of frame types that are annotated is provided as an appendix (i.e., Table 8 and 9).

Given the expertise available on the team and the resources provided through the DFG Collaborative Research Center 991 at the University of Düsseldorf, we are confident that the annotations will be completed on time, and with the highest quality.

Access to copyrighted material: We will provide free access to the copyrighted materials from the WSJ corpus through an agreement with the LDC[16]. We plan to publish and release our annotated dataset (preferably under the Creative Commons license, or as a free resource under a LDC license) by the end of the SemEval task. We hope that this will help us achieve the broader goal of building a gold-standard evaluation benchmark for the unsupervised identification and extraction of coarse semantic frames.

4 The Pilot System

As mentioned earlier, we have developed a pilot system that addresses our proposed task of unsupervised frame induction (whose system description paper is currently under review). The pilot system is based on a hierarchical Bayesian network. In brief, we assume that frames and roles/slots are the latent variables of a probabilistic model. In that model, the probability of a specific frame f with head v, roles $r_1 \dots r_k$ filled by words $w_1 \dots w_k$ and corresponding syntactic dependencies $d_1 \dots d_k$ are given as:

$$p(f) \cdot p(v|f) \prod_{i=1}^{k} p(d_i|f) \cdot \prod_{i=1}^{k} p(r_i|f, d_i) \cdot \prod_{i=1}^{k} p(w_i|r_i).$$

[16] Participants will be provided with a free evaluation license.

Method	pu	ipu	fpu	bcp	bcr	bcf	ari	dst
AllIn1	22.35	100	36.54	9.55	100	17.43	0	36.47
1CpI	100	0.54	1.08	100	0.54	1.08	0.38	N.D.
1CpH	94.38	59.59	73.06	93.54	48.1	63.53	51.29	71.63
Random	24.98	1.99	3.68	14.03	0.94	1.77	0.36	2.69
Pilot	75.07	65.52	69.97	67.77	56.71	61.74	48.43	68.24

Table 6: Clustering data points to frame types: Results from the pilot system and the related baselines. The remaining abbreviations are introduced in Section 2.1 and 2.2.

Method	pu	ipu	fpu	bcp	bcr	bcf	ari	dst
AllIn1	47.64	100	64.53	38.22	100	55.31	0	64.48
1CpI	100	0.18	0.36	100	0.18	0.36	0.04	N.D.
1CpG	92.68	79.62	85.65	87.33	67.97	76.44	66.72	85.36
Random	47.79	4.44	8.12	38.41	3.5	6.41	0.04	7.81
Pilot	90.16	79.04	84.23	83.64	67	74.39	64.39	83.98

Table 7: clustering of syntactic arguments to Semantic Roles.

To estimate the parameters of this model, we use the expectation maximization algorithm and the split-merge clustering technique. Table 6 and Table 7, respectively, list the *best* obtained results from this system for mapping verbs to frame type clusters and syntactic arguments to slot fillers, along with the other baselines listed in Section 2.2 on the current evaluation dataset. The paper that describes the pilot system is currently under review.

5 Task Organizers

- **Marie Candito** is assistant professor at Paris Diderot University, member of the Laboratoire de Linguistique Formelle. Her research interests are in dependency treebanks and dependency parsing, and shallow semantic parsing, focusing on how to better deal with syntax-semantic divergences. She has coordinated the creation of the ASFALDA French FrameNet (http://asfalda.linguist.univ-paris-diderot.fr/frameIndex.xml). email: marie.candito@linguist.univ-paris-diderot.fr.

- **Laura Kallmeyer** is a professor of computational linguistics at Heinrich Heine University Düsseldorf, Germany. She is a renowned expert in language modeling, grammar formalisms, computational semantics and statistical parsing, and she has conducted seminal research in these areas. In 2017, she received an ERC Consolidator Grant for conducting a research project on tree-rewriting grammars, implementation tools for precision grammars and grammar induction and statistical parsing, with a particular focus on a frame-based syntax-semantics interface. Furthermore, Laura Kallmeyer is also the spokesperson of the Collaborative Research Center 991 funded by the DFG on frames as representation structures in language, cognition and science. She is herself involved in the CRC with two scientific projects, one of them on frame induction. email: kallmeyer@phil.hhu .de.

- **Timm Lichte** is a postdoctoral researcher at the at the Collaborative Research Center "The Structure of Representations in Language, Cognition, and Science" located at Heinrich Heine University Düsseldorf, Germany. His research interests include deep syntactic parsing with frame-semantic composition and the identification and modeling of non-literal language, in particular idiomatic multi-word expressions. email: lichte@phil.hhu.de.

- **Alfredo Maldonado** is a postdoctoral researcher in the ADAPT Centre at Trinity College Dublin. His research interests are on computational lexical semantics, terminology and lexicography. In particular, his research focuses on enriching distributional lexical representations (e.g. word embeddings) with linguistic data, the automatic identification of multi-word expressions and terminology and their handling and integration within larger NLP workflows. email: `alfredo.maldonado@adaptcentre.ie`

- **Rainer Osswald** is a senior researcher at the Collaborative Research Center "The Structure of Representations in Language, Cognition, and Science" located at Heinrich Heine University Düsseldorf, Germany. He holds a PhD in Computer Science from the University of Hagen. The focus of his current research is on the formal modeling of the syntax-semantics interface of verb-based constructions by means of decompositional frame semantics. email: `osswald@phil.hhu.de`

- **Miriam R. L. Petruck** received her Ph.D. in Linguistics from the University of California, Berkeley, CA, with the guidance of the late Charles J. Fillmore. A key member of the team developing FrameNet almost since the project's founding (in 1997), Petruck's research interests include semantics, lexical semantics, knowledge base development, grammar and lexis, semantics, Frame Semantics and Construction Grammar, particularly as these linguistic theories support advances in NLU and NLP. She is a frequent invited speaker, lecturing about Frame Semantics, Construction Grammar, and FrameNet internationally; email: `miriamp@icsi.berkeley.edu`.

- **Behrang QasemiZadeh** is a postdoctoral researcher at the Computational Linguistics department of the University of Düsseldorf and a member of the DFG-founded collaborative research center 991, where he investigate the use of data-driven methods for 'hierarchical frame induction'. Behrang's research interests are corpus-based computational linguistics, particularly in applications for lexicography and terminology. Behrang was a co-organizer of SemEval 2018 Task 7 on relation extraction and classification from scientific abstracts. His CV can be accessed in http://pars.ie/cv. email: `zadeh@phil.hhu.de`.

References

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

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

Bagga, A. and Baldwin, B. (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.

Baisa, V., Bradbury, J., Cinkova, S., El Maarouf, I., Kilgarriff, A., and Popescu, O. (2015). Semeval-2015 task 15: A cpa dictionary-entry-building task. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 315–324. Association for Computational Linguistics.

Buitelaar, P. (2005). Ontology learning from text: An overview. *Ontology learning from text: Methods, evaluation and applications*, 123:3–12.

Camacho-Collados, J., Pilehvar, M. T., Collier, N., and Navigli, R. (2017). Semeval-2017 task 2: Multilingual and cross-lingual semantic word similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 15–26.

Cinková, S., Fučíková, E., Šindlerová, J., and Hajič, J. (2014). EngVallex - English valency lexicon. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University.

Damonte, M. and Cohen, S. B. (2017). Cross-lingual abstract meaning representation parsing. *CoRR*, abs/1704.04539.

Das, D., Chen, D., Martins, A. F. T., Schneider, N., and Smith, N. A. (2014). Frame-semantic parsing. *Comput. Linguist.*, 40(1):9–56.

Dohare, S. and Karnick, H. (2017). Text summarization using abstract meaning representation. *CoRR*, abs/1706.01678.

Fillmore, C. J. (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.

Gamerschlag, T., Gerland, D., Osswald, R., and Petersen, W. (2014). *General Introduction*, pages 3–21. Springer International Publishing, Cham.

Girju, R., Nakov, P., Nastase, V., Szpakowicz, S., Turney, P., and Yuret, D. (2007). Semeval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 13–18. Association for Computational Linguistics.

Green, R., Dorr, B. J., and Resnik, P. (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.

Hartmann, S., Kuznetsov, I., Martin, T., and Gurevych, I. (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.

Hendrickx, I., Kim, S. N., Kozareva, Z., Nakov, P., Ó Séaghdha, D., Padó, S., Pennacchiotti, M., Romano, L., and Szpakowicz, S. (2009). Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, DEW '09, pages 94–99, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jurgens, D. and Klapaftis, I. (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.

Jurgens, D., Pilehvar, M. T., and Navigli, R. (2014). Semeval-2014 task 3: Cross-level semantic similarity. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 17–26.

Jurgens, D. A., Turney, P. D., Mohammad, S. M., and Holyoak, K. J. (2012). Semeval-2012 task 2: Measuring degrees of relational similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, pages 356–364, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kshirsagar, M., Thomson, S., Schneider, N., Carbonell, J., Smith, N. A., and Dyer, C. (2015). Frame-semantic role labeling with heterogeneous annotations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 218–224, Beijing, China. Association for Computational Linguistics.

Lang, J. and Lapata, M. (2011). Unsupervised semantic role induction via split-merge clustering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1117–1126, Stroudsburg, PA, USA. Association for Computational Linguistics.

Manandhar, S., Klapaftis, I., Dligach, D., and Pradhan, S. (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.

May, J. and Priyadarshi, J. (2017). Semeval-2017 task 9: Abstract meaning representation parsing and generation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 536–545, Vancouver, Canada. Association for Computational Linguistics.

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

Modi, A., Titov, I., and Klementiev, A. (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.

Navigli, R. and Vannella, D. (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.

Oepen, S., Kuhlmann, M., Miyao, Y., Zeman, D., Cinkova, S., Flickinger, D., Hajic, J., Ivanova, A., and Uresova, Z. (2016). Towards comparability of linguistic graph banks for semantic parsing. In Chair), N. C. C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Pantel, P. and Lin, D. (2002). Document clustering with committees. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, pages 199–206, New York, NY, USA. ACM.

Pennacchiotti, M., De Cao, D., Basili, R., Croce, D., and Roth, M. (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.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850.

Roth, B., Monath, N., Belanger, D., Strubell, E., Verga, P., and McCallum, A. (2015). Building knowledge bases with universal schema: Cold start and slot-filling approaches.

Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, C. R., Baker, C. F., and Scheffczyk, J. (2016). *FrameNet II: Extended Theory and Practice.* ICSI, Berkeley.

Schuster, S. and Manning, C. D. (2016). Enhanced English Universal Dependencies: An improved representation for natural language understanding tasks. In Chair), N. C. C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

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

Surdeanu, M., Johansson, R., Meyers, A., Màrquez, L., and Nivre, J. (2008). The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 159–177. Association for Computational Linguistics.

Swier, R. S. and Stevenson, S. (2004). Unsupervised semantic role labelling. In Lin, D. and Wu, D., editors, *Proceedings of EMNLP 2004*, pages 95–102, Barcelona, Spain. Association for Computational Linguistics.

Appendix A List of annotated frame types

Table 8 lists frame types and statistics related to the annotated instances so far in our data set. Table 9 lists some of the most frequent mappings between verb-lemmas to frame types.

Frame Type	#Inst.	#V
Change_position_on_a_scale	1259	17
Commerce_buy	670	2
Giving	432	7
Cotheme	399	13
Sending	290	5
Awareness	290	6
Activity_start	290	2
Process_start	196	4
Building	172	5
Cogitation	153	4
Cause_to_amalgamate	142	12
Cause_motion	128	10
Placing	121	21
Avoiding	102	8
Assessing	97	3
Choosing	90	3
Judgment_communication	86	14
Scrutiny	77	9
Hiring	62	2
Prohibiting	58	3
Preventing	43	2
Cause_change_of_position_on_a_scale	43	3
Filling	35	12
Thwarting	35	2
Adorning	26	10
Amalgamation	17	3
Seeking	11	3

Table 8: List of Frame types and the frequency of their instances (#Inst.) that are annotated so far. #V denotes the number of different verb lemmas that evoke the frame. Definition for these frames can be found in the FrameNet's repository.

Verb-lemma	**Frame Type**	**Freq.**
buy	Commerce_buy	545
give	Giving	379
fall	Change_position_on_a_scale	305
rise	Change_position_on_a_scale	237
believe	Awareness	225
begin	Activity_start	161
build	Building	150
consider	Cogitation	143
follow	Cotheme	138
begin	Process_start	137
post	Sending	135
drop	Change_position_on_a_scale	120
decline	Change_position_on_a_scale	108
start	Activity_start	103
lead	Cotheme	101
send	Sending	99
purchase	Commerce_buy	91
gain	Change_position_on_a_scale	87
join	Cause_to_amalgamate	86
jump	Change_position_on_a_scale	77
avoid	Avoiding	75
place	Placing	60
value	Assessing	56
hire	Hiring	56
pursue	Cotheme	54
climb	Change_position_on_a_scale	51
choose	Choosing	51
start	Process_start	42
prevent	Preventing	39
conduct	Cotheme	38
soar	Change_position_on_a_scale	36
move	Cause_motion	35
prevent	Thwarting	34
increase	Change_position_on_a_scale	34
increase	Cause_change_of_position_on_a_scale	33
ban	Prohibiting	33
understand	Awareness	30
criticize	Judgment_communication	29
ship	Sending	28
investigate	Scrutiny	28

Table 9: The list of 40 most frequent verb-lemma to frame type mappings.