

Inducing Example-based Semantic Frames from a Massive Amount of Verb Uses

Daisuke Kawahara[†] Daniel W. Peterson[‡] Octavian Popescu[§] Martha Palmer[‡]

[†]Kyoto University, Kyoto, Japan

[‡]University of Colorado at Boulder, Boulder, CO, USA

[§]Fondazione Bruno Kessler, Trento, Italy

dk@i.kyoto-u.ac.jp, {Daniel.W.Peterson, Martha.Palmer}@colorado.edu, popescu@fbk.eu

Abstract

We present an unsupervised method for inducing semantic frames from verb uses in giga-word corpora. Our semantic frames are verb-specific example-based frames that are distinguished according to their senses. We use the Chinese Restaurant Process to automatically induce these frames from a massive amount of verb instances. In our experiments, we acquire broad-coverage semantic frames from two giga-word corpora, the larger comprising 20 billion words. Our experimental results indicate the effectiveness of our approach.

1 Introduction

Semantic frames are indispensable knowledge for semantic analysis or text understanding. In the last decade, semantic frames, such as FrameNet (Baker et al., 1998) and PropBank (Palmer et al., 2005), have been manually elaborated. These resources are effectively exploited in many natural language processing (NLP) tasks, including not only semantic parsing but also machine translation (Boas, 2002), information extraction (Surdeanu et al., 2003), question answering (Narayanan and Harabagiu, 2004), paraphrase acquisition (Ellsworth and Janin, 2007) and recognition of textual entailment (Burchardt and Frank, 2006).

There have been many attempts to automatically acquire frame knowledge from raw corpora with the goal of either adding frequency information to an existing resource or of inducing similar frames for other languages. Most of these approaches, however, focus on syntactic frames, i.e., subcategorization frames (e.g., (Manning, 1993; Briscoe and Carroll, 1997; Korhonen et al., 2006; Lippincott et al., 2012; Reichart and Korhonen, 2013)). Since subcategorization frames represent

argument patterns of verbs and are purely syntactic, expressions that have the same subcategorization frame can have different meanings (e.g., metaphors). Semantics-oriented NLP applications based on frames, such as paraphrase acquisition and machine translation, require consistency in the meaning of each frame, and thus these subcategorization frames are not suitable for these semantic tasks.

Recently, there have been a few studies on automatically acquiring semantic frames (Materna, 2012; Materna, 2013). Materna induced semantic frames (called LDA-Frames) from triples of (subject, verb, object) in the British National Corpus (BNC) based on Latent Dirichlet Allocation (LDA) and the Dirichlet Process. LDA-Frames capture limited linguistic phenomena of these triples, and are defined across verbs based on probabilistic topic distributions.

This paper presents a method for automatically building verb-specific semantic frames from a large raw corpus. Our semantic frames are verb-specific like PropBank and semantically distinguished. A frame has several syntactic case slots, each of which consists of words that are eligible to fill the slot. For example, let us show three semantic frames of the verb “observe”:¹

observe:1

nsubj:{we, author, ...} dobj:{effect, result, ...}
prep_in:{study, case, ...} ...

observe:2

nsubj:{teacher, we, ...} dobj:{child, student, ...}
prep_in:{classroom, school, ...} ...

observe:3

nsubj:{child, people, ...} dobj:{bird, animal, ...}
prep_at:{range, time, ...} ...

¹In this paper, we use the dependency relation names of the Stanford collapsed dependencies (de Marneffe et al., 2006) as the notations of case slots. For instance, “nsubj” means a nominal subject, “dobj” means a direct object, “iboj” means an indirect object, “ccomp” means a clausal complement and “prep_*” means a preposition.

Frequencies, which are not shown in the above examples, are attached to each semantic frame, case slot and word, and can be effectively exploited for the applications of these semantic frames. The frequencies of words in each case slot become good sources of selectional preferences.

Our novel contributions are summarized as follows:

- induction of semantic frames based on the Chinese Restaurant Process (Aldous, 1985) from only automatic parses of a web-scale corpus,
- exploitation of the assumption of one sense per collocation (Yarowsky, 1993) to make the computation feasible,
- providing broad-coverage knowledge for selectional preferences, and
- evaluating induced semantic frames by using an existing annotated corpus with verb classes.

2 Related Work

The most closely related work to our semantic frames are LDA-Frames, which are probabilistic semantic frames automatically induced from a raw corpus (Materna, 2012; Materna, 2013). He used a model based on LDA and the Dirichlet Process to cluster verb instances of a triple (subject, verb, object) to produce semantic frames and slots. Both of these are represented as a probabilistic distribution of words across verbs. He applied this method to the BNC and acquired 427 frames and 144 slots (Materna, 2013). These frames are over-generalized across verbs and might be difficult to provide with fine-grained selectional preferences. In addition, Grenager and Manning (2006) proposed a method for inducing PropBank-style frames from Stanford typed dependencies extracted from raw corpora. Although these frames are based on typed dependencies and more semantic than subcategorization frames, they are not distinguished in terms of the senses of words filling a case slot.

There are hand-crafted semantic frames in the lexicons of FrameNet (Baker et al., 1998) and PropBank (Palmer et al., 2005). Corpus Pattern Analysis (CPA) frames (Hanks, 2012) are another manually created repository of patterns for verbs. Each pattern represents a prototypical word usage as extracted by lexicographers from the BNC. Creating CPA is time consuming, but our proposed

method may be employed to assist in the creation of this type of resource, as shown in Section 4.4.

Our task can be regarded as clustering of verb instances. In this respect, the models of Parisien and Stevenson are related to our method (Parisien and Stevenson, 2009; Parisien and Stevenson, 2010). Parisien and Stevenson (2009) proposed a Dirichlet Process model for clustering usages of the verb “get.” Later, Parisien and Stevenson (2010) proposed a Hierarchical Dirichlet Process model for jointly clustering argument structures (i.e., subcategorization frames) and verb classes. However, their argument structures are not semantic but syntactic, and also they did not evaluate the resulting frames. There have also been related approaches to clustering verb types (Vlachos et al., 2009; Sun and Korhonen, 2009; Falk et al., 2012; Reichart and Korhonen, 2013). These methods induce verb clusters in which multiple verbs participate, and do not consider the polysemy of verbs. Our objective is different from theirs.

Another line of related work is unsupervised semantic parsing or semantic role labeling (Poon and Domingos, 2009; Lang and Lapata, 2010; Lang and Lapata, 2011a; Lang and Lapata, 2011b; Titov and Klementiev, 2011; Titov and Klementiev, 2012). These approaches basically cluster predicates and their arguments to distinguish predicate senses and semantic roles of arguments. Modi et al. (2012) extended the model of Titov and Klementiev (2012) to jointly induce semantic roles and frames using the Chinese Restaurant Process, which is also used in our approach. However, they did not aim at building a lexicon of semantic frames, but at distinguishing verbs that have different senses in a relatively small annotated corpus. Applying this method to a large corpus could produce a frame lexicon, but its scalability would be a big problem.

For other languages than English, Kawahara and Kurohashi (2006a) proposed a method for automatically compiling Japanese semantic frames from a large web corpus. They applied conventional agglomerative clustering to predicate-argument structures using word/frame similarity based on a manually-crafted thesaurus. Since Japanese is head-final and has case-marking postpositions, it seems easier to build semantic frames with it than with other languages such as English. They also achieved an improvement in dependency parsing and predicate-argument structure

analysis by using their resulting frames (Kawahara and Kurohashi, 2006b).

3 Method for Inducing Semantic Frames

Our objective is to automatically induce verb-specific example-based semantic frames. Each semantic frame consists of a partial set of syntactic slots: *nsubj*, *dobj*, *iobj*, *ccomp* and *prep_**. Each slot consists of words with frequencies, which could provide broad-coverage selectional preferences.

Frames for a verb should be semantically distinguished. That is to say, each frame should consist of predicate-argument structures that have consistent usages or meanings.

Our procedure to automatically generate semantic frames from verb usages is as follows:

1. apply dependency parsing to a raw corpus and extract predicate-argument structures for each verb from the automatic parses,
2. merge the predicate-argument structures that have presumably the same meaning based on the assumption of one sense per collocation to get a set of initial frames, and
3. apply clustering to the initial frames based on the Chinese Restaurant Process to produce the final semantic frames.

Each of these steps is described in the following sections in detail.

3.1 Extracting Predicate-argument Structures from a Raw Corpus

We first apply dependency parsing to a large raw corpus. We use the Stanford parser with Stanford dependencies (de Marneffe et al., 2006).² Collapsed dependencies are adopted to directly extract prepositional phrases.

Then, we extract predicate-argument structures from the dependency parses. Dependents that have the following dependency relations to a verb are extracted as arguments:

```
nsubj, xsubj, dobj, iobj, ccomp, xcomp,
prep_*
```

Here, we do not distinguish adjuncts from arguments. All extracted dependents of a verb are handled as arguments. This distinction is left for future work, but this will be performed using slot

²<http://nlp.stanford.edu/software/lex-parser.shtml>

Sentences:

They observed the effects of ...
 This statistical ability to observe an effect ...
 We did not observe a residual effect of ...
 He could observe the results at the same time ...
 My first opportunity to observe the results of ...
 You can observe beautiful birds ...
 Children may then observe birds ...
 ⋮

Predicate-argument structures:

```
nsubj:they observe dobj:effect
observe dobj:effect
nsubj:we observe dobj:effect
nsubj:he observe dobj:result prep_at:time
observe dobj:result
nsubj:you observe dobj:bird
nsubj:child observe dobj:bird
⋮
```

Initial frames:

```
nsubj:{they, we, ...} observe dobj:{effect}
nsubj:{he, ...} observe dobj:{result} prep_at:{time}
nsubj:{you, child, ...} observe dobj:{bird}
⋮
```

Figure 1: Examples of predicate-argument structures and initial frames for the verb “observe.”

frequencies in the applications of semantic frames or the method proposed by Abend and Rappoport (2010).

We apply the following processes to extracted predicate-argument structures:

- A verb and an argument are lemmatized, and only the head of an argument is preserved for compound nouns.
- Phrasal verbs are also distinguished from non-phrasal verbs. For example, “look up” has independent frames from “look.”
- The passive voice of a verb is distinguished from the active voice, and thus these have independent frames. Passive voice is detected using the part-of-speech tag “VBN” (past participle). The alignment between frames of active and passive voices will be done after the induction of frames using the model of Sasano et al. (2013) in the future.
- “xcomp” (open clausal complement) is renamed to “ccomp” (clausal complement) and “xsubj” (controlling subject) is renamed to “nsubj” (nominal subject). This is because

these usages as predicate-argument structures are not different.

- A capitalized argument with the part-of speech “NNP” (singular proper noun) or “NNPS” (plural proper noun) is generalized to $\langle \text{name} \rangle$. Similarly, an argument of “ccomp” is generalized to $\langle \text{comp} \rangle$ since the content of a clausal complement is not important.

Extracted predicate-argument structures are collected for each verb and the subsequent processes are applied to the predicate-argument structures of each verb. Figure 1 shows examples of predicate-argument structures for “observe.”

3.2 Constructing Initial Frames from Predicate-argument Structures

A straightforward way to produce semantic frames is to cluster the extracted predicate-argument structures directly. Since our objective is to compile broad-coverage semantic frames, a massive amount of predicate-argument structures should be fed into the clustering. It would take prohibitive computational costs to conduct the sampling procedure, which is described in the next section.

To make the computation feasible, we merge the predicate-argument structures that have the same or similar meaning to get initial frames. These initial frames are the input of the subsequent clustering process. For this merge, we assume one sense per collocation (Yarowsky, 1993) for predicate-argument structures.

For each predicate-argument structure of a verb, we couple the verb and an argument to make a unit for sense disambiguation. We select an argument in the following order by considering the degree of effect on the verb sense:³

dobj, ccomp, nsubj, prep_*, iobj.

This selection of a predominant argument order above is justified by relative comparisons of the discriminative power of the different slots for CPA frames (Popescu, 2013). If a predicate-argument structure does not have any of the above slots, it is discarded.

Then, the predicate-argument structures that have the same verb and argument pair (slot and

³If a predicate-argument structure has multiple prepositional phrases, one of them is randomly selected.

word, e.g., “dobj:effect”) are merged into an initial frame (Figure 1). After this process, we discard minor initial frames that occur fewer than 10 times.

For example, we have 732,292 instances (predicate-argument structures) for the verb “observe” in the web corpus that is used in our experiment (its details are described in Section 4.1). As the result of this merging process, we obtain 6,530 initial frames, which become an input for the clustering. This means that this process accelerates the speed of clustering more than 100 times.

The precision of this process will be evaluated in Section 4.3.

3.3 Clustering using Chinese Restaurant Process

We cluster initial frames for each verb to produce final semantic frames using the Chinese Restaurant Process (Aldous, 1985). We regard each initial frame as an instance in the usual clustering of the Chinese Restaurant Process.

We calculate the posterior probability of a semantic frame f_j given an initial frame v_i as follows:

$$P(f_j|v_i) \propto \begin{cases} \frac{n(f_j)}{N+\alpha} \cdot P(v_i|f_j) & f_j \neq \text{new} \\ \frac{\alpha}{N+\alpha} \cdot P(v_i|f_j) & f_j = \text{new}, \end{cases} \quad (1)$$

where N is the number of initial frames for the target verb and $n(f_j)$ is the current number of initial frames assigned to the semantic frame f_j . α is a hyper-parameter that determines how likely it is for a new semantic frame to be created. In this equation, the first term is the Dirichlet process prior and the second term is the likelihood of v_i .

$P(v_i|f_j)$ is defined based on the Dirichlet-Multinomial distribution as follows:

$$P(v_i|f_j) = \prod_{w \in V} P(w|f_j)^{\text{count}(v_i, w)}, \quad (2)$$

where V is the vocabulary in all case slots cooccurring with the verb. It is distinguished by the case slot, and thus consists of pairs of slots and words, e.g., “nsubj:child” and “dobj:bird.” $\text{count}(v_i, w)$ is the number of w in the initial frame v_i .

$P(w|f_j)$ is defined as follows:

$$P(w|f_j) = \frac{\text{count}(f_j, w) + \beta}{\sum_{t \in V} \text{count}(f_j, t) + |V| \cdot \beta}, \quad (3)$$

where $count(f_j, w)$ is the current number of w in the frame f_j , and β is a hyper-parameter of Dirichlet distribution. For a new semantic frame, this probability is uniform ($1/|V|$).

We use Gibbs sampling to realize this clustering.

4 Experiments and Evaluations

4.1 Experimental Settings

We use two kinds of large-scale corpora: a web corpus and the English Gigaword corpus.

To prepare a web corpus, we first crawled the web. We extracted sentences from each web page that seems to be written in English based on the encoding information. Then, we selected sentences that consist of at most 40 words, and removed duplicated sentences. From this process, we obtained a corpus of one billion sentences, totaling approximately 20 billion words. We focused on verbs whose frequency was more than 1,000. There were 19,649 verbs, including phrasal verbs, and separating passive and active constructions. We extracted 2,032,774,982 predicate-argument structures.

We also used the English Gigaword corpus (LDC2011T07; English Gigaword Fifth Edition) to induce semantic frames. This corpus consists of approximately 180 million sentences, which totaling four billion words. There were 7,356 verbs after applying the same frequency threshold as the web corpus. We extracted 423,778,278 predicate-argument structures from this corpus.

We set the hyper-parameters α in (1) and β in (3) to 1.0. The frame assignments for all the components were initialized randomly. We took 100 samples for each initial frame and selected the frame assignment that has the highest probability. These parameters were determined according to a preliminary experiment to manually examine the quality of resulting frames.

4.2 Experimental Results

We executed the per-verb clustering tasks on a PC cluster. It finished within a few hours for most verbs, but it took a couple of days for very frequent verbs, such as “get” and “say.” The clustering produced an average number of semantic frames per verb of 15.2 for the web corpus and 18.5 for the Gigaword corpus. Examples of induced semantic frames from the web corpus are shown in Table 1.

	slot	instances
observe:1	nsubj	i:5850, we:5201, he:3796, you:3669, ...
	doobj	what:7091, people:2272, this:2262, ...
	prep_in	way:254, world:204, life:194, ...
	⋮	
observe:2	nsubj	we:11135, you:1321, i:1317, ...
	doobj	change:5091, difference:2719, ...
	prep_in	study:622, case:382, cell:362, ...
	⋮	
observe:3	nsubj	student:3921, i:2240, we:2174, ...
	doobj	child:2323, class:2184, student:2025, ...
	prep_in	classroom:555, action:509, ...
	⋮	
accept:1	nsubj	we:44833, i:6873, order:4051, ...
	doobj	card:28835, payment:22569, ...
	prep_for	payment:1166, convenience:1147, ...
	⋮	
accept:2	nsubj	i:10568, we:9300, you:5106, ...
	doobj	that:14180, this:12061, it:7756, ...
	prep_as	part:1879, fact:1085, truth:926, ...
	⋮	
accept:3	nsubj	people:7459, he:6696, we:5515, ...
	doobj	christ:13766, jesus:6528, it:5612, ...
	prep_as	savior:5591, lord:597, one:469, ...
	⋮	

Table 1: Examples of resulting frames for the verb “observe” and “accept” induced from the web corpus. The number following an instance word represents its frequency.

4.3 Evaluation of Induced Semantic Frames

We evaluate precision and coverage of induced semantic frames. To measure the precision of induced semantic frames, we adopt the purity metric, which is usually used to evaluate clustering results. However, the problem is that it is impossible to assign gold-standard classes to the huge number of instances. To automatically measure the purity of the induced semantic frames, we make use of the SemLink corpus (Loper et al., 2007), in which VerbNet classes (Kipper-Schuler, 2005) and PropBank/FrameNet frames are assigned to each instance. We make a test set that contains 157 polysemous verbs that occur 10 or more times in the SemLink corpus (sections 02-21 of the Wall Street Journal). We first add these instances to the instances from a raw corpus and apply clustering to these merged instances. Then, we compare the induced semantic frames of the SemLink instances with their gold-standard classes. We adopt VerbNet classes and PropBank frames as gold-standard classes.

For each group of verb-specific semantic frames, we measure the purity of the frames as the percentage of SemLink instances belonging to the majority gold class in their respective cluster. Let

		PU		CO		F ₁	
		Mac	Mic	Mac	Mic	Mac	Mic
against VerbNet	One frame	0.799	0.802	0.917	0.952	0.854	0.870
	Initial frames	0.985	0.982	0.755	0.812	0.855	0.889
	Induced sem frames	0.900	0.901	0.886	0.928	0.893	0.914
against PropBank	One frame	0.901	0.872	↑	↑	0.909	0.910
	Initial frames	0.994	0.993	↑	↑	0.858	0.893
	Induced sem frames	0.965	0.949	↑	↑	0.924	0.939

Table 2: Evaluation results of semantic frames from the web corpus against VerbNet classes and PropBank frames. “Mac” means a macro average and “Mic” means a micro average.

		PU		CO		F ₁	
		Mac	Mic	Mac	Mic	Mac	Mic
against VerbNet	One frame	0.799	0.804	0.855	0.920	0.826	0.858
	Initial frames	0.985	0.981	0.666	0.758	0.795	0.855
	Induced sem frames	0.916	0.909	0.796	0.880	0.852	0.894
against PropBank	One frame	0.901	0.874	↑	↑	0.877	0.896
	Initial frames	0.994	0.993	↑	↑	0.798	0.859
	Induced sem frames	0.968	0.953	↑	↑	0.874	0.915

Table 3: Evaluation results of semantic frames from the Gigaword corpus against VerbNet classes and PropBank frames. “Mac” means a macro average and “Mic” means a micro average.

N denote the total number of SemLink instances of the target verb, G_j the set of instances belonging to the j -th gold class and F_i the set of instances belonging to the i -th frame. The purity (PU) can then be written as follows:

$$PU = \frac{1}{N} \sum_j \max_i |G_j \cap F_i|. \quad (4)$$

For example, a frame of the verb “observe” contains 11 SemLink instances, and eight out of them belong to the class SAY-37.7, which is the majority class among these 11 instances. PU is calculated by summing up such counts over all the frames of this verb.

Usually, inverse purity or collocation is used to measure the recall of normal clustering tasks. However, these recall measures do not fit our task. This is because it is not a real error to have similar separate frames. Instead, we want to avoid having so many frames that we cannot provide broad-coverage selectional preferences due to sparsity. To judge this aspect, we measure coverage.

The coverage (CO) measures to what extent predicate-argument structures of the target verb in a test set are included in one of frames of the verb. We use the predicate-argument structures of the above 157 verbs from the SemLink corpus, which are the same ones used in the evaluation of PU. We judge a predicate-argument structure as correct if all of its argument words (of the target slot

described in Section 3.1) are included in the corresponding slot of a frame. If the clustering gets better, the value of CO will get higher, because merging instances by clustering alleviates data sparsity.

These per-verb scores are aggregated into an overall score by averaging over all verbs. We use two ways of averaging: a macro average and a micro average. The macro average is a simple average of scores for individual verbs. The micro average is obtained by weighting the scores for individual verbs proportional to the number of instances for that verb. Finally, we use the harmonic mean (F₁) of purity and coverage as a single measure of clustering quality.

For comparison, we adopt the following two baseline methods:

One frame a frame into which all the instances for a verb are merged

Initial frames the initial frames without clustering (described in Section 3.2)

Table 2 and Table 3 list evaluation results for semantic frames induced from the web corpus and the Gigaword corpus, respectively.⁴ Note that CO does not consider gold-standard classes, and thus the values of CO are the same for the VerbNet

⁴We did not adopt inverse purity, but its values for the induced semantic frames range from 0.42 to 0.49.

and PropBank evaluations. The induced frames outperformed the two baseline methods in terms of F_1 in most cases. While the coverage of the web frames was higher than that of the Gigaword frames, as expected, the purity of the web frames was slightly lower than that of the Gigaword frames. This degradation might be caused by the noise in the web corpus.

The purity of the initial frames was around 98%-99%, which means that there were few cases that the one-sense-per-collocation assumption was violated.

Modi et al. (2012) reported a purity of 77.9% for the assignment of FrameNet frames to the FrameNet corpus. We also conducted the above purity evaluation against FrameNet frames for 140 verbs.⁵ We obtained a macro average of 92.9% and a micro average of 89.2% for the web frames, and a macro average of 93.2% and a micro average of 89.8% for the Gigaword frames. It is difficult to directly compare these results with Modi et al. (2012), but our frame assignments seem to have higher accuracy.

4.4 Evaluation against CPA Frames

Corpus Pattern Analysis (CPA) is a technique for linking word usage to prototypical syntagmatic patterns.⁶ The resource was built manually by investigating examples in the BNC, and the set of corpus examples used to induce each pattern is given. For example, the following three patterns describe the usage of the verb “accommodate.”

[Human 1] accommodate [Human 2]
 [Building] accommodate [Eventuality]
 [Human] accommodate [Self] to [Eventuality]

In this paper, we use CPA to evaluate the quality of the automatically induced frames. By comparing the induced frames to CPA patterns, we can evaluate the correctness and relevance of this approach from a human point of view. To do that, we associate semantic features to the set of words in each slot in the frames, using SUMO (Niles and Pease, 2001). For example, take the following frame for the verb “accomplish”:

accomplish:1

nsubj: {you, leader, employee, ...}
 dobj: {developing, progress, objective, ...}.

⁵Since FrameNet frames are not assigned to all the verbs of SemLink, the number of verbs is different from the evaluations against VerbNet and PropBank.

⁶<http://deb.fi.muni.cz/pdev/>

	all	K-means
Entropy (E)	0.790	0.516
Recovery Rate (RC)	0.347	0.630
Purity (P)	0.462	0.696

Table 4: CPA Evaluation.

Using SUMO, we map this frame to the following:
 nsubj: [Human]
 dobj: [SubjectiveAssessmentAttribute],
 which corresponds to pattern 3 for “accomplish” in CPA.

We also associate SUMO attributes to the CPA patterns with more than 10 examples (716 verbs). There are many patterns of SUMO attributes for any CPA frame or induced frame, since each filler word in a particular slot can have more than one SUMO attribute. We filter out the non-discriminative SUMO attributes following the technique described in Popescu (2013). Using this, we obtain SUMO attributes for both CPA clusters and induced frames, and we can use the standard entropy-based measures to evaluate the match between the two types of patterns: E — entropy, RC — recovery rate, and P — purity (Li et al., 2004):

$$E = \sum_{j=1}^K \frac{m_j}{m} \cdot e_j, \quad RC = 1 - \sum_{j,i=1}^{K,L} \frac{p_{ij}}{m_i}, \quad (5)$$

$$P = \sum_{j=1}^K \frac{m_j}{m} \cdot p_j, \quad p_j = \max_i p_{ij}, \quad (6)$$

$$e_j = \sum_{i=1}^L p_{ij} \log_2 p_{ij}, \quad p_{ij} = \frac{m_{ij}}{m_i}, \quad (7)$$

where m_j is the number of induced frames corresponding to topic j , m_{ij} is the number of induced frames in cluster j and annotated with the CPA pattern i , m is the total number of induced frames, L is the number of CPA patterns, and K is the number of induced frames.

We also consider a K-means clustering process, with K set as 2 or 3 depending on the number of SUMO-attributed patterns. The K-means evaluation is carried out considering only the centroid of the cluster, which corresponds to the prototypical induced semantic frame with SUMO attributes. We compute E , RC and P using formulae (5) - (7) for each verb and then compute the macro average, considering all the frames and only the K-means centroids, respectively. The results for the induced web frames are displayed in Table 4.

The evaluation method presented here overcomes some of the drawbacks of the previous approaches (Materna, 2012; Materna, 2013). First, we did not limit the evaluation to the most frequent patterns. Second, the mapping was carried out automatically and not by hand. The results above compare favorably with the previous approaches, especially considering that no filtering procedures were applied to the induced frames. We anticipate that the results based on the prototypical induced frames with SUMO attributes would be competitive. Our post-analysis revealed that the entropy can be lowered further if an automatic filtering based on frequencies is applied.

4.5 Evaluation of the Quality of Selectional Preferences

We also investigated the quality of selectional preferences within the induced semantic frames. The only publicly available test data for selectional preferences, to our knowledge, is from Chambers and Jurafsky (2010). This data consists of quadruples (verb, relation, word, confounder) and does not contain their context.⁷

A typical way for using our semantic frames is to select an appropriate frame for an input sentence and judge the eligibility of the word uses against the selected frame. However, due to the lack of context for the above data, it is difficult to select a corresponding semantic frame for a test quadruple and thus the induced semantic frames cannot be naturally applied to this data. To investigate the potential for selectional preferences of the semantic frames, we approximately match a quadruple with each of the semantic frames of the verb and select the frame that has the highest probability as follows:

$$P(w) = \max_i P(w|v, rel, f_i), \quad (8)$$

where w is the word or confounder, v is the verb, rel is the relation and f_i is a semantic frame. By comparing the probabilities of the word and the confounder, we select either of them according to the higher probability. For tie breaking in the case that no frames are found for the verb or both the word and confounder are not found in the case slot, we randomly select either of them in the same way as Chambers and Jurafsky (2010).

We use the “neighbor frequency” set, which is the most difficult among the three sets included

⁷A document ID of the English Gigaword corpus is available, but it is difficult to recover the context of each instance from this information.

in the data. It contains 6,767 quadruples and the relations consist of three classes: subject, object and preposition, which has no distinction of actual prepositions. To link these relations with our case slots, we manually aligned the subject with the *nsubj* (nominal subject) slot, the object with the *dobj* (direct object) slot and the preposition with *prep_** (all the prepositions) slots. For the preposition relation, we choose the highest probability among all the preposition slots in a frame. To match the generalized $\langle name \rangle$ with the word in a quadruple, we change the word to $\langle name \rangle$ if it is capitalized and not a capitalized personal pronoun.

Our semantic frames from the Gigaword corpus achieved an accuracy of 81.7%⁸ and those from the web corpus achieved an accuracy of 80.2%. This slight deterioration seems to come from the noise in the web corpus. The best performance in Chambers and Jurafsky (2010) is 81.7% on this “neighbor frequency” set, which was achieved by conditional probabilities with the Erk (2007)’s smoothing method calculated from the English Gigaword corpus. Our approach for selectional preferences does not use smoothing like Erk (2007), but it achieved equivalent performance to the previous work. If we applied our semantic frames to a verb instance with its context, a more precise judgment of selectional preferences would be possible with appropriate frame selection.

5 Conclusion

This paper has described an unsupervised method for inducing semantic frames from instances of each verb in giga-word corpora. This method is clustering based on the Chinese Restaurant Process. The resulting frame data are open to the public and also can be searched by inputting a verb via our web interface.⁹

As applications of the resulting frames, we plan to integrate them into syntactic parsing, semantic role labeling and verb sense disambiguation. For instance, Kawahara and Kurohashi (2006b) improved accuracy of dependency parsing based on Japanese semantic frames automatically induced from a large raw corpus. It is valuable and promising to apply our semantic frames to these NLP tasks.

⁸Since the dataset was created from the NYT 2001 portion of the English Gigaword Corpus, we built semantic frames again from the Gigaword corpus except this part.

⁹<http://nlp.ist.i.kyoto-u.ac.jp/member/kawahara/cf/crp.en/>

Acknowledgments

This work was supported by Kyoto University John Mung Program and JST CREST. We gratefully acknowledge the support of the National Science Foundation Grant NSF 1116782 - RI: Small: A Bayesian Approach to Dynamic Lexical Resources for Flexible Language Processing. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

- Omri Abend and Ari Rappoport. 2010. Fully unsupervised core-adjunct argument classification. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 226–236.
- David Aldous. 1985. Exchangeability and related topics. *École d'Été de Probabilités de Saint-Flour XIII —1983*, pages 1–198.
- Collin Baker, Charles J. Fillmore, and John Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages 86–90.
- Hans C. Boas. 2002. Bilingual framenet dictionaries for machine translation. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pages 1364–1371.
- Ted Briscoe and John Carroll. 1997. Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, pages 356–363.
- Aljoscha Burchardt and Anette Frank. 2006. Approximating textual entailment with LFG and FrameNet frames. In *Proceedings of the 2nd PASCAL Recognizing Textual Entailment Workshop*, pages 92–97.
- Nathanael Chambers and Daniel Jurafsky. 2010. Improving the use of pseudo-words for evaluating selectional preferences. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 445–453.
- Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 449–454.
- Michael Ellsworth and Adam Janin. 2007. Mutaphrase: Paraphrasing with framenet. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 143–150.
- Katrin Erk. 2007. A simple, similarity-based model for selectional preferences. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 216–223.
- Ingrid Falk, Claire Gardent, and Jean-Charles Lamirel. 2012. Classifying french verbs using french and english lexical resources. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 854–863.
- Trond Grenager and Christopher D. Manning. 2006. Unsupervised discovery of a statistical verb lexicon. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 1–8.
- Patrick Hanks. 2012. How people use words to make meanings: Semantic types meet valencies. *Input, Process and Product: Developments in Teaching and Language Corpora*, pages 54–69.
- Daisuke Kawahara and Sadao Kurohashi. 2006a. Case frame compilation from the web using high-performance computing. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 1344–1347.
- Daisuke Kawahara and Sadao Kurohashi. 2006b. A fully-lexicalized probabilistic model for Japanese syntactic and case structure analysis. In *Proceedings of the Human Language Technology Conference of the NAACL*, pages 176–183.
- Karin Kipper-Schuler. 2005. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania.
- Anna Korhonen, Yuval Krymolowski, and Ted Briscoe. 2006. A large subcategorization lexicon for natural language processing applications. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 345–352.
- Joel Lang and Mirella Lapata. 2010. Unsupervised induction of semantic roles. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 939–947.
- Joel Lang and Mirella Lapata. 2011a. Unsupervised semantic role induction via split-merge clustering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1117–1126.
- Joel Lang and Mirella Lapata. 2011b. Unsupervised semantic role induction with graph partitioning. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1320–1331.
- Tao Li, Sheng Ma, and Mitsunori Ogihara. 2004. Entropy-based criterion in categorical clustering. In *Proceedings of the 21st International Conference on Machine Learning*, volume 4, pages 536–543.

- Thomas Lippincott, Anna Korhonen, and Diarmuid Ó Séaghdha. 2012. Learning syntactic verb frames using graphical models. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 420–429.
- Edward Loper, Szu-Ting Yi, and Martha Palmer. 2007. Combining lexical resources: mapping between PropBank and VerbNet. In *Proceedings of the 7th International Workshop on Computational Linguistics*.
- Christopher Manning. 1993. Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 235–242.
- Jiří Materna. 2012. LDA-Frames: An unsupervised approach to generating semantic frames. In Alexander Gelbukh, editor, *Proceedings of the 13th International Conference CICLing 2012, Part I*, volume 7181 of *Lecture Notes in Computer Science*, pages 376–387. Springer Berlin / Heidelberg.
- Jiří Materna. 2013. Parameter estimation for LDA-Frames. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 482–486.
- 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.
- Srini Narayanan and Sanda Harabagiu. 2004. Question answering based on semantic structures. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 693–701.
- Ian Niles and Adam Pease. 2001. Towards a standard upper ontology. In *Proceedings of the International Conference on Formal Ontology in Information Systems*, pages 2–9.
- Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.
- Christopher Parisien and Suzanne Stevenson. 2009. Modelling the acquisition of verb polysemy in children. In *Proceedings of the CogSci2009 Workshop on Distributional Semantics beyond Concrete Concepts*, pages 17–22.
- Christopher Parisien and Suzanne Stevenson. 2010. Learning verb alternations in a usage-based Bayesian model. In *Proceedings of the 32nd annual meeting of the Cognitive Science Society*.
- Hoifung Poon and Pedro Domingos. 2009. Unsupervised semantic parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1–10.
- Octavian Popescu. 2013. Learning corpus patterns using finite state automata. In *Proceedings of the 10th International Conference on Computational Semantics*, pages 191–203.
- Roi Reichart and Anna Korhonen. 2013. Improved lexical acquisition through DPP-based verb clustering. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 862–872.
- Ryohei Sasano, Daisuke Kawahara, Sadao Kurohashi, and Manabu Okumura. 2013. Automatic knowledge acquisition for case alternation between the passive and active voices in Japanese. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1213–1223.
- Lin Sun and Anna Korhonen. 2009. Improving verb clustering with automatically acquired selectional preferences. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 638–647.
- Mihai Surdeanu, Sanda Harabagiu, John Williams, and Paul Aarseth. 2003. Using predicate-argument structures for information extraction. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 8–15.
- Ivan Titov and Alexandre Klementiev. 2011. A Bayesian model for unsupervised semantic parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1445–1455.
- Ivan Titov and Alexandre Klementiev. 2012. A Bayesian approach to unsupervised semantic role induction. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 12–22.
- Andreas Vlachos, Anna Korhonen, and Zoubin Ghahramani. 2009. Unsupervised and constrained dirichlet process mixture models for verb clustering. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 74–82.
- David Yarowsky. 1993. One sense per collocation. In *Proceedings of the Workshop on Human Language Technology*, pages 266–271.