

# <sup>[NS]</sup><sub>S</sub> provisional title:]Leveraging Additional Resources for Frame-Semantic Role Labeling

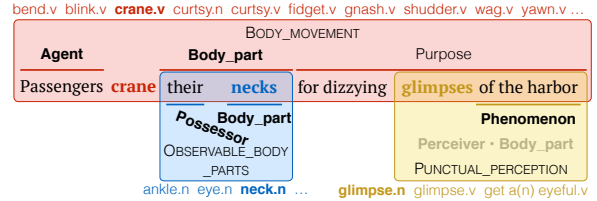
## Abstract

The high cost of semantic structure annotation is a major obstacle to automating semantic analysis with broad coverage. The fully annotated datasets that exist are often small, hindering the robustness of models trained on them. However, low-resource tasks may benefit from exploiting *partially* annotated data, as well as data with *different* (but related) forms of annotation, for additional training data or features. This paper considers the argument identification and classification subtask of frame-semantic parsing, which to date has relied exclusively upon full-text annotations in the FrameNet resource. <sup>[NS]</sup><sub>S</sub> is this true? I think Dipanjan used exemplars for the latent variable frame ID model, but he hasn’t used them at all for arg ID, right?] We augment supervised learning with additional “indirect” training data and features so as to leverage additional resources internal and external to FrameNet (e.g., PropBank). <sup>[NS]</sup><sub>S</sub> result]

## 1 Introduction

<sup>[NS]</sup><sub>S</sub> sparseness is a challenge for many computational semantics tasks]

Frame-semantic parsing (Das et al., 2014) is a case in point. This is the task of automating the rich linguistic structure analyses of the FrameNet lexicon and corpus (Baker et al., 1998).<sup>1</sup> FrameNet represents kinds of events, scenarios, and relationships with an inventory of **frames** (<sup>[NS]</sup><sub>S</sub> examples]). Each frame is associated with lexical **predicates** (verbs, nouns, adjectives, and adverbs) capable of evoking the scenario, and a set of **roles** (or **frame elements**) called to mind in order to understand



**Figure 1:** Example sentence from FrameNet full-text annotation. 3 frames and their arguments are shown: BODY\_MOVEMENT is evoked by *crane*, OBSERVABLE\_BODY\_PARTS by *necks*, and PUNCTUAL\_PERCEPTION by *glimpse*. (Further, *harbor* is annotated as evoking the LOCALE\_BY\_USE frame and doubles as its sole argument.) Horizontal lines representing argument spans are labeled with role names.

the scenario. These roles may be implicit, but are frequently realized linguistically in the same sentence as the predicate. Given a sentence, frame-semantic parsing is the task of mapping tokens in the sentences to evoked frames, and for each evoked frame, finding and labeling its **argument** phrases with roles. An example appears in figure 1; it will be explained in detail in §2.2.

FrameNet 1.5 defines a structured hierarchy of over 1,000 frames associated with <sup>[NS]</sup><sub>S</sub> # English lexical predicates, and also provides annotations for <sup>[NS]</sup><sub>S</sub> # targets annotated total] attestations of these frames/predicates in corpora, annotated in context with their arguments. In FrameNet 1.5, a rather small number of sentences—<sup>[NS]</sup><sub>S</sub> #, comprising <sup>[NS]</sup><sub>S</sub> # words—are provided with **full-text** annotations, i.e. the sentence has been analyzed for all available frames. But a full <sup>[NS]</sup><sub>S</sub> # % of sentences in FrameNet—the lexicographic **exemplars**—are annotated for only one frame per sentence, and have thus far not been exploited successfully for frame-semantic parsing. Here, we seek to leverage these exemplar sentences as well as the (type-level) hierarchical structure of the FrameNet lexicon.

In this paper, we address the **argument identification** subtask of finding and labeling arguments given a predicate in context and the frame

<sup>1</sup><http://framenet.icsi.berkeley.edu/>

| A0         |       | A1    |       | AM-PRP     |                                 |
|------------|-------|-------|-------|------------|---------------------------------|
| Passengers | crane | their | necks | for        | dizzying glimpses of the harbor |
|            |       |       |       | A0         |                                 |
|            |       |       |       | A1         |                                 |
|            |       |       |       | dizzy-v-01 |                                 |

**Figure 2:** Ideal PropBank annotations for verbs. Though PB uses lexical frames rather than deep frames, there are clear similarities to the FrameNet annotations in figure 1.

it evokes. This is a form of semantic role labeling (SRL). [<sub>S</sub><sup>T</sup> cite Gildea, Roth, Toutanova, etc.] Notably, another resource, **PropBank** (Kingsbury and Palmer, 2002), has been widely used for SRL (Palmer et al., 2010). PropBank annotations capture shallower lexical frames and arguments; additionally, PropBank provides [<sub>S</sub><sup>NS</sup> millions?] of words of fully annotated English sentences (annotation is much less expensive, but also potentially less valuable, because of the shallower representation). Despite a number of differences in the representations and annotation conventions, for many predicates FrameNet and PropBank are really quite similar: figure 1 and figure 2 show this for the verb *crane*. To get the best of both worlds, we aim to tap into PropBank’s vast resources as indirect token-level supervision for FrameNet-style analysis. We hypothesize that PropBank analyses can serve as a weak signal for the FrameNet SRL task, either by heuristically transforming PropBank annotations into FrameNet annotations to augment the training data, or by preprocessing sentences with a PropBank SRL system to obtain new features for FrameNet argument identification.

Our experiments expand the *training data* and/or the *feature space* of supervised argument identification in order to integrate evidence from all of these sources into SEMAFOR (Das et al., 2014), the leading open-source frame-semantic parser for English.<sup>2</sup> The results show that some of these sources of evidence succeed at boosting argument identification performance.[<sub>S</sub><sup>NS</sup> SOTA (without constraints)?]

## 2 Resources

[<sub>S</sub><sup>NS</sup> incl. data analysis of differences from full-text]  
 [<sub>S</sub><sup>NS</sup> w/in each mention: genre/overall vocabulary; coverage and distributions of predicates, FE labels; oracle coverage of FT test]

<sup>2</sup><http://www.ark.cs.cmu.edu/SEMAFOR/>

## 2.1 The FrameNet Lexicon

Each frame in the Berkeley FrameNet lexicon is intended to represent a gestalt scene. The frame definition includes: a descriptive name; a set of **core roles** representing participants and props that are crucial to understanding the scene; a set of **non-core roles** such as circumstantial information (time, place, manner, purpose, etc.); an English textual description of the scene and how its roles relate to one another; and a set of English predicates that can evoke the scene. For example, the BODY\_MOVEMENT frame has **Agent** and **Body\_part** as its core roles; the frame description states, “This frame contains words for motions or actions an **Agent** performs using some part of his/her body.” Lexical entries in this frame include verbs such as bend, blink, crane, and curtsy, plus the noun use of curtsy.

[<sub>S</sub><sup>NS</sup> hierarchy. which relations do we care about—anything besides Inheritance?]

## 2.2 Full-text Annotations

Contemporary frame-semantic parsers are trained and evaluated on the **full-text (FT)**<sup>3</sup> portion of the FrameNet corpus. This consists of documents for which annotators made an effort to assign frames and arguments to as many words as possible. Figure 1 gives an example sentence from the FT portion of the corpus. It has 4 frame annotations. BODY\_MOVEMENT, as described in the previous section, is evoked by *crane*, and 3 of its roles are filled by overt arguments: the 2 core roles (**Agent**, **Body\_part**) happen to be filled by noun phrases (*passengers*, *their necks*), while the non-core role **Purpose** is filled by a prepositional phrase adjunct (*for dizzying glimpses of the harbor*). The frame defines 19 additional non-core roles, none of which have an argument in the example. In frame semantics, non-core roles are considered to be *conceptually* optional; core roles may or may not be *syntactically* optional, but if not locally specified they are expected to be available from context, or else implicit. For example, PUNCTUAL\_PERCEPTION—evoked in this case by *glimpses*—is annotated as missing 2 of its core roles. A human listener would resolve the identity of the **Perceiver** from the wider context and the **Body\_part** from world knowledge.

In some cases, FT annotation involved creating a new frame or adding a new lexical unit to an ex-

<sup>3</sup>Though these were *annotated* and the document level, and train/dev/test splits are by document, the frame-semantic parsing is currently restricted to the sentence level.

isting frame. In other cases, words that in principle should be considered to evoke a frame were left unannotated because they did not match any existing lexical units. This was the case for *passengers* and *dizzying* in figure 1.

Genres, sizes

Annotation density: (proportion of tokens evoking a frame. breakdown by POS?)

## 2.3 Exemplars

[<sup>NS</sup><sub>S</sub> how different from FT]

## 2.4 PropBank

[<sup>NS</sup><sub>S</sub> how different from FN]

## 2.5 SemLink

[<sup>NS</sup><sub>S</sub> limitations!]

## 2.6 Illinois SRL system

## 3 Learning from multiple domains and representations

We experiment with several techniques for modifying the model-fitting portion of the argument identification model’s local learning objective (training data, features). All of our experiments use the same form of regularization, condition on the same frame predictions[<sup>NS</sup><sub>S</sub> not oracle, right? these are SEMAFOR’s current frame ID model, so not state of the art?] and syntactic preprocessing[<sup>NS</sup><sub>S</sub> does this match Dipanjan’s latest experiments?], and use beam search with [<sup>NS</sup><sub>S</sub> hyperparam value] for joint decoding of arguments.<sup>4</sup>

[<sup>NS</sup><sub>S</sub> Domain adaptation/multitask learning techniques]

### 3.1 Augmenting the Training Data

### 3.2 Frustratingly Easy

### 3.3 2-stage

### 3.4 Type-level hierarchy features

[<sup>M</sup><sub>K</sub> Types of relations and the number of relations of each type used: Inheritance and Subframe. How many types of relations?]

Frames in FrameNet are connected to each other by relations such as inheritance, temporal ordering, causality. For instance, the frame ROBBERY inherits from the more abstract frame *Committing\_crime*, the frame FALL\_ASLEEP is preceded by the frame

<sup>4</sup>[<sup>NS</sup><sub>S</sub> recent work has improved upon global decoding techniques; we expect such improvements to be complementary to the gains due to the local model reported here]

BEING\_AWAKE. The roles of related frames have also been mapped to indicate the correspondence between them: ROBBERY.**Perpetrator** is mapped to COMMITTING\_CRIME.**Perpetrator**, which in turn maps to MISDEED.**Wrongdoer**. Frames and roles that are far apart in this hierarchy are less related than say neighbours. This hierarchy can be exploited to share information across related roles, thereby benefiting the roles that have few annotations [<sup>M</sup><sub>K</sub> say something about a greater variety of contexts is available for each role]. A simple mechanism to share information is via shared model parameters between related roles. Towards this, we define two types of hierarchical features: (1) ‘siblings’, where for every feature  $f_i$  that fires for an argument  $a$ , we add a feature which is the conjunction:  $(f_i \wedge \text{parent.frame} \wedge \text{parent.role} \wedge I_{\text{hier}})$  where  $\text{parent.frame} = \text{parent}(\text{frame}(a))$ .  $I_{\text{hier}}$  is an indicator to distinguish this feature from the regular conjunction features that use frame names and roles. (2) ‘parents’, where the addendum is the feature:  $(f_i \wedge \text{frame} \wedge \text{frame.role} \wedge I_{\text{hier}})$ . We use only the ‘Inheritance’ and ‘Subframe’ relations between the roles, of which there are 4138 and 589 respectively.

## 4 Experiments

[<sup>NS</sup><sub>S</sub> tuning regularizer for all experiments]

### 4.1 Baseline

Our baseline system is SEMAFOR (Das et al., 2014). SEMAFOR treats the argument identification task as a structured prediction problem. It uses a linearly parametrized model that scores each candidate span given a frame element[<sup>S</sup><sub>T</sub> this should probably be talked about earlier?]:

$$\text{score}_{\mathbf{w}}(y | x) = \mathbf{w}^T \mathbf{f}(x, y) \quad (1)$$

Automatic syntactic parses [<sup>S</sup><sub>T</sub> from where? are we using the MST-stacked dep parses?] are used to narrow the set of candidate spans considered, and as input to feature extraction.

During training, each frame element is treated as an independent multiclass logistic regression instance, with the set of candidate spans (including the NULL span) as its output space. But at test time, it chooses a joint assignment of all arguments that maximizes probability under the model, while satisfying the following constraints:

1. an argument may be assigned to at most one span [<sup>S</sup><sub>T</sub> theta criterion, Chomsky], and

2. spans of realized arguments must not overlap.

Beam search, with a beam size of 100, is used to choose the maximum joint assignment with no overlapping arguments. <sup>[S<sup>NS</sup> does beam search require normalizing to probabilities?]</sup>

We have made several modifications to SEMAFOR’s training that do not affect performance, but do speed up experiments:

- We use squared structured hinge loss (defined below) instead of multiclass logistic regression. Using hinge loss, there is no longer a need to calculate a partition function. Gradients, and hence parameters, are sparser than in logistic regression, allowing us to use a sparse vector implementation.
- We use the online optimization method AdaDelta (Zeiler, 2012) with minibatches <sup>[S<sup>NS</sup> minibatch size?]</sup>, instead of the batch method L-BFGS (Liu and Nocedal, 1989).
- We use elastic net ( $\ell_1 + \ell_2$ ) regularization. Adding  $\ell_1$  also has the effect of keeping the parameter vector sparse. <sup>[S<sup>NS</sup> what hyperparameter(s)? are  $\ell_1$  and  $\ell_2$  tuned separately?]</sup>

We use these changes for all systems, including the baseline. While performance is not significantly affected <sup>[M<sup>M</sup> P/R/F numbers are in the table (see rows 1 and 2). Times: 12 hrs 9 mins for the earlier algorithm to converge. That took 290 iterations. An equivalent number of iterations took 45 minutes for the new system. Running till convergence (about 700 iterations) took 82 minutes, thus giving a speed-up of  $\approx 9X$ . The P/R/F on exemplars improves (row 2); I believe it is due to the regularization in Sam’s objective.]</sup>, these changes enabled us to run more experiments with the larger exemplar dataset and expanded feature space.

The details of squared hinge loss are as follows. Let  $(x^{(i)}, y^{(i)})$  be the  $i^{th}$  training example. Then the structured hinge loss on the  $i^{th}$  example is given by:

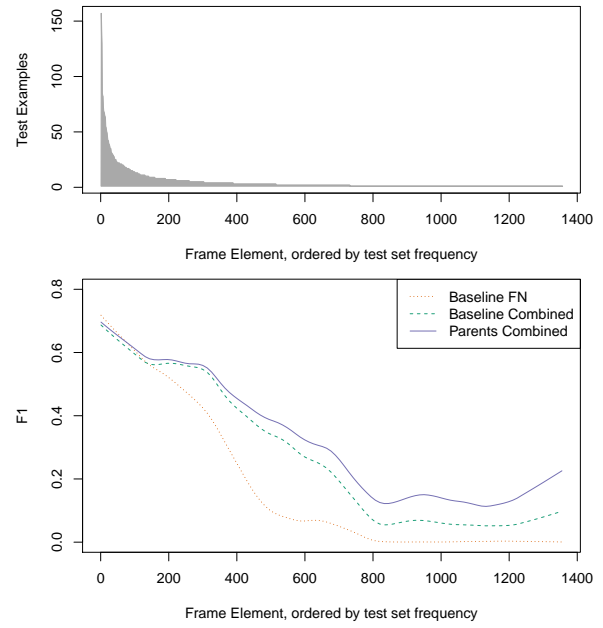
$$\text{Hinge}_{\mathbf{w}}(i) = \max_y \{ \mathbf{w}^T \mathbf{f}(x^{(i)}, y) + \text{cost}(y, y^{(i)}) \} - \mathbf{w}^T \mathbf{f}(x^{(i)}, y^{(i)})$$

and squared hinge loss is:

$$\text{SquaredHinge}_{\mathbf{w}}(i) = \text{Hinge}_{\mathbf{w}}(i)^2. \quad (2)$$

In our experiments, we use  $\text{cost}(y, y^{(i)}) = \mathbf{1}\{y \neq y^{(i)}\}$  <sup>[S<sup>S</sup> define]<sup>5</sup></sup>.

<sup>5</sup>We experimented with recall-oriented training, where



**Figure 3:** Count and  $F_1$  for each frame element appearing in the test set.  $F_1$  values have been smoothed with loess, with a smoothing parameter of 0.2.

<sup>[S<sup>NS</sup> same FN 1.5 splits as Dipanjan]</sup>

<sup>[S<sup>NS</sup> preprocessing issues: removing duplicate sentences, merging adjacent split args in exemplars, OntoNotes PropBank preprocessing (NLTK), token-level SemLink details (such as filtering out sentences without mappable annotations; copy from WS paper)]</sup>

## 4.2 Evaluation

<sup>[S<sup>NS</sup> main eval: FT test; new eval: exemplars]</sup>

## 4.3 Results

<sup>[S<sup>NS</sup> results table without hierarchy features]</sup>

<sup>[S<sup>NS</sup> hierarchy features: which ones work best (decided on baseline), how do they improve best result so far]</sup>

<sup>[S<sup>NS</sup> Sam’s curves on per-FE  $F_1$ ]</sup>

<sup>[S<sup>NS</sup> comparison to prior work (baseline, best result). args+frames score vs. args only]</sup>

<sup>[S<sup>NS</sup> discussion throughout]</sup>

## 5 Related Work

<sup>[S<sup>NS</sup> Dipanjan’s other papers; mention other PB SRL work?; anything using SemLink or combining resources for SRL?]</sup>

errors of omission are assigned a higher cost, but found that while recall increased, overall  $F_1$  went down <sup>[S<sup>NS</sup> or: failed to improve?]</sup>.

| Additional Resource | Training Configuration (Features)   | Full-Text |       |       | Exemplars |       |       |
|---------------------|---|-----------|-------|-------|-----------|-------|-------|
|                     |   | P         | R     | $F_1$ | P         | R     | $F_1$ |
| (Baseline)          | FT (Basic)  | 66.03     | 53.79 | 59.29 | 64.90     | 33.60 | 44.27 |
| FN Hierarchy        | FT (siblings)<br>FT (siblings+parents)  |           |       |       |           |       |       |
| Exemplars           | Exemplars $\xrightarrow{\text{guide}}$ FT<br>FT+Exemplars (Basic)<br>FT+Exemplars (EasyAdapt) |           |       |       |           |       |       |
| SemLink             | SemLink $\xrightarrow{\text{guide}}$ FT<br>FT+SemLink   |           |       |       |           |       |       |
| PB-SRL              | FT (PB-SRL)   |           |       |       |           |       |       |

**Table 1:** Results on two test sets: Baseline vs. individual other resources. Precision, recall, and  $F_1$  are given as percentages.

FT+Exemplars (Hier: siblings+parents)  
 FT+Exemplars (PB-SRL)  
 FT+Exemplars (PB-SRL, Hier: siblings+parents)

**Table 2:** Combining best techniques across resources [S<sup>NS</sup> TODO]

|  | Test on FN |         |         | Test on Exemplars |         |         |
|--|------------|---------|---------|-------------------|---------|---------|
|  | P          | R       | $F_1$   | P                 | R       | $F_1$   |
| Semafor baseline (Ddas' model)           | 0.6603     | 0.5379  | 0.5929  | 0.64933           | 0.33582 | 0.44269 |
| Semafor baseline (Sam's code)            | 0.65569    | 0.53820 | 0.59116 | 0.6263            | 0.3765  | 0.4703  |
| baseline trained on combined             | 0.66061    | 0.58234 | 0.61901 | 0.75443           | 0.65107 | 0.69895 |
| trained only on exemplars                | 0.61084    | 0.49049 | 0.54409 | 0.77010           | 0.65958 | 0.71057 |
| trained on FN + exemplars guide features | 0.65241    | 0.55960 | 0.60245 | 0.67709           | 0.48076 | 0.56228 |
| frust <sup>†</sup> on combined           | 0.65702    | 0.59043 | 0.62195 | 0.73876           | 0.61397 | 0.67061 |
| siblings <sup>‡</sup> on FN              | 0.67244    | 0.54763 | 0.60365 | 0.64815           | 0.39088 | 0.48766 |
| siblings, trained on combined            | 0.65991    | 0.60406 | 0.63075 | 0.76140           | 0.67713 | 0.71679 |
| parents <sup>*</sup> on FN               | 0.67672    | 0.52790 | 0.59312 | 0.65250           | 0.38184 | 0.48176 |
| parents, trained on combined             | 0.65920    | 0.60382 | 0.63029 | 0.76143           | 0.68317 | 0.72018 |
| trained on semlink                       | 0.44433    | 0.10533 | 0.17029 | 0.48119           | 0.12498 | 0.19842 |
| trained on FN + semlink guide features   | 0.64671    | 0.54533 | 0.59171 | 0.60951           | 0.38922 | 0.47507 |
| trained on FN+SemLink                    | 0.655      | 0.3776  | 0.4791  | 0.57148           | 0.20780 | 0.30478 |
| with SRL augmented spans and features    | 0.70550    | 0.53178 | 0.60644 |                   |         |         |

**combined:** FN + Exemplars training data

†: feature augmentation from the frustratingly easy DA paper

‡: for every feature  $f_i$  that fires for an argument  $a$ , fire an additional feature which is the conjunction:  $(f_i \wedge \text{parent.frame} \wedge \text{parent.role} \wedge I_{\text{hier}})$  where  $\text{parent.frame} = \text{parent}(\text{frame}(a))$ . The parent's frame and role are obtained from the FN hierarchy.  $I_{\text{hier}}$  is an indicator to distinguish this feature from the regular conjunction features that use frame names and roles.

\*: fire the siblings feature<sup>‡</sup> and an additional feature:  $(f_i \wedge \text{frame} \wedge \text{frame.role} \wedge I_{\text{hier}})$

[<sup>NS</sup><sub>S</sub> multitask learning?]

## 6 Conclusion

[<sup>NS</sup><sub>S</sub> overall findings]

[<sup>NS</sup><sub>S</sub> future work: testing ground for improvements to PB and SemLink; automatic mappings between resources]

## References

- Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proc. of COLING-ACL*, pages 86–90. Montreal, Quebec, Canada.
- Dipanjan Das, Desai Chen, André F. T. Martins, Nathan Schneider, and Noah A. Smith. 2014. Frame-semantic parsing. *Computational Linguistics*, 40(1):9–56.
- Paul Kingsbury and Martha Palmer. 2002. From TreeBank to PropBank. In *Proc. of LREC*, pages 1989–1993. Las Palmas, Canary Islands.
- Dong C. Liu and Jorge Nocedal. 1989. On the Limited Memory BFGS Method for Large Scale Optimization. *Math. Program.*, 45(3):503–528.
- Martha Palmer, Daniel Gildea, and Nianwen Xue. 2010. *Semantic Role Labeling*. Number 6 in Synthesis Lectures on Human Language Technologies. Morgan & Claypool, San Rafael, CA.
- Matthew Zeiler. 2012. ADADELTA: An Adaptive Learning Rate Method. *CoRR*, abs/1212.5701. URL <http://dblp.uni-trier.de/rec/bibtex/journals/corr/abs-1212-5701>.