Probabilistic Frame-Semantic Parsing

Dipanjan Das Nathan Schneider Desai Chen Noah A. Smith
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{dipanjan@cs,nschneid@cs,desaic@andrew,nasmith@cs}.cmu.edu

Abstract

This paper contributes a formalization of frame-semantic parsing as a structure prediction problem and describes an implemented parser that transforms an English sentence into a frame-semantic representation. It finds words that evoke FrameNet frames, selects frames for them, and locates the arguments for each frame. The system uses two feature-based, discriminative probabilistic (log-linear) models, one with latent variables to permit disambiguation of new predicate words. The parser is demonstrated to significantly outperform previously published results.

1 Introduction

FrameNet (Fillmore et al., 2003) is a rich linguistic resource containing considerable information about lexical and predicate-argument semantics in English. Grounded in the theory of frame semantics (Fillmore, 1982), it suggests—but does not formally define—a semantic representation that blends word-sense disambiguation and semantic role labeling.

In this paper, we present a computational and statistical model for frame-semantic parsing, the problem of extracting from text semantic predicate-argument structures such as those shown in Fig. 1. We aim to predict a frame-semantic representation as a *structure*, not as a pipeline of classifiers. We use a probabilistic framework that cleanly integrates the FrameNet lexicon and (currently very limited) available training data. Although our models often involve strong independence assumptions, the probabilistic framework we adopt is highly amenable to future extension through new features, relaxed independence assumptions, and semisupervised learning. Some novel aspects of our current approach include a latent-variable model that permits disambiguation of words not in the FrameNet lexicon, a unified model for finding and labeling arguments,

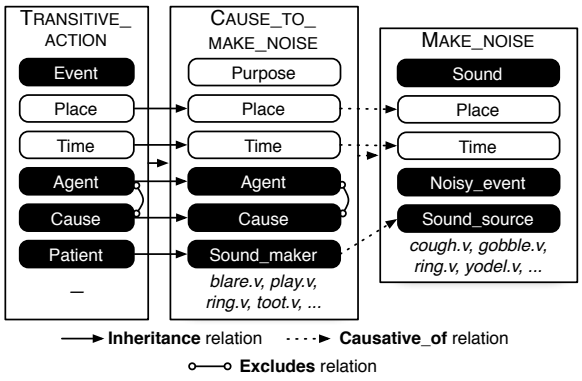

Figure 2. Partial illustration of frames, roles, and LUs related to the CAUSE_TO_MAKE_NOISE frame, from the FrameNet lexicon. "Core" roles are filled ovals. 8 additional roles of CAUSE_TO_MAKE_NOISE are not shown.

and a precision-boosting constraint that forbids arguments of the same predicate to overlap. Our parser achieves the best published results to date on the SemEval'07 FrameNet task (Baker et al., 2007).

2 Resources and Task

We consider frame-semantic parsing resources.

2.1 FrameNet Lexicon

The FrameNet lexicon is a taxonomy of manually identified general-purpose **frames** for English.[1] Listed in the lexicon with each frame are several lemmas (with part of speech) that can denote the frame or some aspect of it—these are called **lexical units** (LUs). In a sentence, word or phrase tokens that evoke a frame are known as **targets**. The set of LUs listed for a frame in FrameNet may not be exhaustive; we may see a target in new data that does not correspond to an LU for the frame it evokes. Each frame definition also includes a set of frame elements, or **roles**, corresponding to different aspects of the concept represented by the frame, such as participants, props, and attributes. We use the term **ar-**

[1] Like the SemEval'07 participants, we used FrameNet v. 1.3 (http://framenet.icsi.berkeley.edu).

Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 948–956,
Los Angeles, California, June 2010.

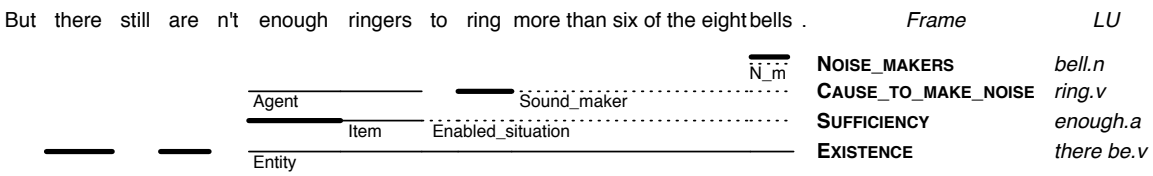

Figure 1. A sentence from PropBank and the SemEval'07 training data, and a partial depiction of gold FrameNet annotations. Each frame is a row below the sentence (ordered for readability). Thick lines indicate targets that evoke frames; thin solid/dotted lines with labels indicate arguments. "N_m" under *bells* is short for the Noise_maker role of the NOISE_MAKERS frame. The last row indicates that *there... are* is a discontinuous target. In PropBank, the verb *ring* is the only annotated predicate for this sentence, and it is not related to other predicates with similar meanings.

FRAMENET LEXICON V. 1.3		
lexical entries	**exemplars**	
	counts	*coverage*
8379 LUs	139K sentences, 3.1M words	70% LUs
795 frames	1 frame annotation / sentence	63% frames
7124 roles	285K overt arguments	56% roles

Table 1. Snapshot of lexicon entries and exemplar sentences. Coverage indicates the fraction of types attested in at least one exemplar.

FULL-TEXT ANNOTATIONS	*SemEval'07 data*		
	train	**dev**	**test**
Size	*(words sentences documents)*		
all	43.3K 1.7K 22	6.3K 251 4	2.8K 120 3
ANC (travel)	3.9K 154 2	.8K 32 1	1.3K 67 1
NTI (bureaucratic)	32.2K 1.2K 15	5.5K 219 3	1.5K 53 2
PropBank (news)	7.3K 325 5	0 0 0	0 0 0
Annotations	*(frames/word overt arguments/word)*		
all	0.23 0.39	0.22 0.37	0.37 0.65
Coverage of lexicon	*(% frames % roles % LUs)*		
all	64.1 27.4 21.0	34.0 10.2 7.3	29.3 7.7 4.9
Out-of-lexicon types	*(frames roles LUs)*		
all	14 69 71	2 4 2	39 99 189
Out-of-lexicon tokens	*(% frames % roles % LUs)*		
all	0.7 0.9 1.1	1.0 0.4 0.2	9.8 11.2 25.3

Table 2. Snapshot of the SemEval'07 annotated data.

gument to refer to a sequence of word tokens annotated as filling a frame role. Fig. 1 shows an example sentence from the training data with annotated targets, LUs, frames, and role-argument pairs. The FrameNet lexicon also provides information about relations between frames and between roles (e.g., INHERITANCE). Fig. 2 shows a subset of the relations between three frames and their roles.

Accompanying most frame definitions in the FrameNet lexicon is a set of lexicographic **exemplar sentences** (primarily from the British National Corpus) annotated for that frame. Typically chosen to illustrate variation in argument realization patterns for the frame in question, these sentences only contain annotations for a single frame. We found that using exemplar sentences directly to train our models hurt performance as evaluated on SemEval'07 data, even though the number of exemplar sentences is an order of magnitude larger than the number of sentences in our training set (§2.2). This is presumably because the exemplars are neither representative as a sample nor similar to the test data. Instead, we make use of these exemplars in features (§4.2).

2.2 Data

Our training, development, and test sets consist of documents annotated with frame-semantic structures for the SemEval'07 task, which we refer to collectively as the **SemEval'07 data**.[2] For the most part, the frames and roles used in annotating these documents were defined in the FrameNet lexicon, but there are some exceptions for which the annotators defined supplementary frames and roles; these are included in the possible output of our parser.

Table 2 provides a snapshot of the SemEval'07 data. We randomly selected three documents from the original SemEval training data to create a development set for tuning model hyperparameters. Notice that the test set contains more annotations per word, both in terms of frames and arguments. Moreover, there are many more out-of-lexicon frame, role, and LU types in the test set than in the training set. This inconsistency in the data results in poor recall scores for all models trained on the given data split, a problem we have not sought to address here.

[2] http://framenet.icsi.berkeley.edu/semeval/FSSE.html

Preprocessing. We preprocess sentences in our dataset with a standard set of annotations: POS tags from MXPOST (Ratnaparkhi, 1996) and dependency parses from the MST parser (McDonald et al., 2005) since manual syntactic parses are not available for most of the FrameNet-annotated documents. We used WordNet (Fellbaum, 1998) for lemmatization. We also labeled each verb in the data as having ACTIVE or PASSIVE voice, using code from the SRL system described by Johansson and Nugues (2008).

2.3 Task and Evaluation

Automatic annotations of frame-semantic structure can be broken into three parts: (1) *targets*, the words or phrases that evoke frames; (2) the *frame type*, defined in the lexicon, evoked by each target; and (3) the *arguments*, or spans of words that serve to fill roles defined by each evoked frame. These correspond to the three subtasks in our parser, each described and evaluated in turn: target identification (§3), frame identification (§4, not unlike word-sense disambiguation), and argument identification (§5, not unlike semantic role labeling).

The standard evaluation script from the SemEval'07 shared task calculates precision, recall, and F_1-measure for frames and arguments; it also provides a score that gives partial credit for hypothesizing a frame related to the correct one. We present precision, recall, and F_1-measure microaveraged across the test documents, report *labels-only* matching scores (spans must match exactly), and do not use named entity labels. More details can be found in Baker et al. (2007). For our experiments, statistical significance is measured using a reimplementation of Dan Bikel's randomized parsing evaluation comparator.[3]

2.4 Baseline

A strong baseline for frame-semantic parsing is the system presented by Johansson and Nugues (2007, hereafter J&N'07), the best system in the SemEval'07 shared task. For frame identification, they used an SVM classifier to disambiguate frames for known frame-evoking words. They used WordNet synsets to extend the vocabulary of frame-evoking words to cover unknown words, and then used a collection of separate SVM classifiers—one for each frame—to predict a single evoked frame for each occurrence of a word in the extended set.

J&N'07 modeled the argument identification problem by dividing it into two tasks: first, they classified candidate spans as to whether they were arguments or not; then they assigned roles to those that were identified as arguments. Both phases used SVMs. Thus, their formulation of the problem involves a multitude of classifiers—whereas ours uses two log-linear models, each with a single set of weights, to find a full frame-semantic parse.

TARGET IDENTIFICATION	P	R	F_1
Our technique (§3)	**89.92**	**70.79**	**79.21**
Baseline: J&N'07	*87.87*	*67.11*	*76.10*

Table 3. Target identification results for our system and the baseline. Scores in bold denote significant improvements over the baseline ($p < 0.05$).

3 Target Identification

Target identification is the problem of deciding which word tokens (or word token sequences) evoke frames in a given sentence. In other semantic role labeling schemes (e.g. PropBank), simple part-of-speech criteria typically distinguish predicates from non-predicates. But in frame semantics, verbs, nouns, adjectives, and even prepositions can evoke frames under certain conditions. One complication is that semantically-impoverished **support predicates** (such as *make* in *make a request*) do not evoke frames in the context of a frame-evoking, syntactially-dependent noun (*request*). Furthermore, only temporal, locative, and directional senses of prepositions evoke frames.

We found that, because the test set is more completely annotated—that is, it boasts far more frames per token than the training data (see Table 2)—learned models did not generalize well and achieved poor test recall. Instead, we followed J&N'07 in using a small set of rules to identify targets.

For a span to be a candidate target, it must appear (up to morphological variation) as a target in the training data or the lexicon. We consider multiword targets,[4] unlike J&N'07 (though we do not consider

[3] `http://www.cis.upenn.edu/~dbikel/software.html#comparator`

[4] There are 629 multiword LUs in the lexicon, and they correspond to 4.8% of the targets in the training set; among them are *screw up*.V, *shoot the breeze*.V, and *weapon of mass de-*

FRAME IDENTIFICATION (§4)	*targets*	exact frame matching *P*	*R*	F_1	partial frame matching *P*	*R*	F_1
Frame identification (oracle targets)	$*$	60.21	60.21	60.21	74.21	74.21	74.21
Frame identification (predicted targets)	auto §3	**69.75**	**54.91**	**61.44**	**77.51**	**61.03**	**68.29**
Baseline: J&N'07	*auto*	*66.22*	*50.57*	*57.34*	*73.86*	*56.41*	*63.97*

Table 4. Frame identification results. Precision, recall, and F_1 were evaluated under exact and partial frame matching; see §2.3. Bold indicates statistically significant results with respect to the baseline ($p < 0.05$).

discontinuous targets). Using rules from §3.1.1 of J&N'07, we further prune the list, with two modifications: we prune *all* prepositions, including locative, temporal, and directional ones, but do not prune support verbs. This is a conservative approach; our automatic target identifier will never propose a target that was not seen in the training data or FrameNet.

Results. Table 3 shows results on target identification; our system gains 3 F_1 points over the baseline.

4 Frame Identification

Given targets, the parser next identifies their frames.

4.1 Lexical units

FrameNet specifies a great deal of structural information both within and among frames. For frame identification we make use of frame-evoking **lexical units**, the (lemmatized and POS-tagged) words and phrases listed in the lexicon as referring to specific frames. For example, listed with the BRAGGING frame are 10 LUs, including *boast*.N, *boast*.V, *boastful*.A, *brag*.V, and *braggart*.N. Of course, due to polysemy and homonymy, the same LU may be associated with multiple frames; for example, *gobble*.V is listed under both the INGESTION and MAKE_NOISE frames. All targets in the exemplar sentences, and most in our training and test data, correspond to known LUs (see Table 2).

To incorporate frame-evoking expressions found in the training data but not the lexicon—and to avoid the possibility of lemmatization errors—our frame identification model will incorporate, via a latent variable, features based directly on exemplar and training **targets** rather than LUs. Let $\mathcal{L}$ be the set of (unlemmatized and automatically POS-tagged) targets found in the exemplar sentences of the lexicon and/or the sentences in our training set. Let $\mathcal{L}_f \subseteq \mathcal{L}$ be the subset of these targets annotated as evoking a particular frame f. Let $\mathcal{L}^l$ and $\mathcal{L}^l_f$ denote the lemmatized versions of $\mathcal{L}$ and $\mathcal{L}_f$ respectively. Then, we write *boasted*.VBD $\in \mathcal{L}_{\text{BRAGGING}}$ and *boast*.VBD $\in \mathcal{L}^l_{\text{BRAGGING}}$ to indicate that this inflected verb *boasted* and its lemma *boast* have been seen to evoke the BRAGGING frame. Significantly, however, another target, such as *toot your own horn*, might be used in other data to evoke this frame. We thus face the additional hurdle of predicting frames for unknown words.

The SemEval annotators created 47 new frames not present in the lexicon, out of which 14 belonged to our training set. We considered these with the 795 frames in the lexicon when parsing new data. Predicting new frames is a challenge not yet attempted to our knowledge (including here). Note that the scoring metric (§2.3) gives partial credit for *related* frames (e.g., a more general frame from the lexicon).

4.2 Model

For a given sentence $\mathbf{x}$ with frame-evoking targets $\mathbf{t}$, let t_i denote the ith target (a word sequence). Let t^l_i denote its lemma. We seek a list $\mathbf{f} = \langle f_1, \ldots, f_m \rangle$ of frames, one per target. In our model, the set of candidate frames for t_i is defined to include every frame f such that $t^l_i \in \mathcal{L}^l_f$—or if $t^l_i \notin \mathcal{L}^l$, then every known frame (the latter condition applies for 4.7% of the gold targets in the development set). In both cases, we let $\mathcal{F}_i$ be the set of candidate frames for the ith target in $\mathbf{x}$.

To allow frame identification for targets whose lemmas were seen in neither the exemplars nor the training data, our model includes an additional variable, ℓ_i. This variable ranges over the seen targets in $\mathcal{L}_{f_i}$, which can be thought of as **prototypes** for the expression of the frame. Importantly, frames are *predicted*, but prototypes are summed over via the latent variable. The prediction rule requires a probabilistic model over frames for a target:

$$f_i \leftarrow \text{argmax}_{f \in \mathcal{F}_i} \sum_{\ell \in \mathcal{L}_f} p(f, \ell \mid t_i, \mathbf{x}) \quad (1)$$

struction.N. In the SemEval'07 training data, there are just 99 discontinuous multiword targets (1% of all targets).

We adopt a conditional log-linear model: for $f \in \mathcal{F}_i$ and $\ell \in \mathcal{L}_f$, $p_{\boldsymbol{\theta}}(f, \ell \mid t_i, \mathbf{x}) =$

$$\frac{\exp \boldsymbol{\theta}^\top \mathbf{g}(f, \ell, t_i, \mathbf{x})}{\sum_{f' \in \mathcal{F}_i} \sum_{\ell' \in \mathcal{L}_{f'}} \exp \boldsymbol{\theta}^\top \mathbf{g}(f', \ell', t_i, \mathbf{x})} \quad (2)$$

where $\boldsymbol{\theta}$ are the model weights, and $\mathbf{g}$ is a vector-valued feature function. This discriminative formulation is very flexible, allowing for a variety of (possibly overlapping) features; e.g., a feature might relate a frame type to a prototype, represent a lexical-semantic relationship between a prototype and a target, or encode part of the syntax of the sentence.

Previous work has exploited WordNet for better coverage during frame identification (Johansson and Nugues, 2007; Burchardt et al., 2005, e.g., by expanding the set of targets using synsets), and others have sought to extend the lexicon itself (see §6). We differ in our use of a latent variable to incorporate lexical-semantic *features* in a discriminative model, relating known lexical units to unknown words that may evoke frames. Here we are able to take advantage of the large inventory of partially-annotated exemplar sentences.

Note that this model makes a strong independence assumption: each frame is predicted independently of all others in the document. In this way the model is similar to J&N'07. However, ours is a single conditional model that shares features and weights across all targets, frames, and prototypes, whereas the approach of J&N'07 consists of many separately trained models. Moreover, our model is unique in that it uses a latent variable to smooth over frames for unknown or ambiguous LUs.

Frame identification features depend on the preprocessed sentence $\mathbf{x}$, the prototype ℓ and its WordNet lexical-semantic relationship with the target t_i, and of course the frame f. Our model instantiates 662,020 binary features; see Das et al. (2010).

4.3 Training

Given the training subset of the SemEval'07 data, which is of the form $\left\langle \langle \mathbf{x}^{(j)}, \mathbf{t}^{(j)}, \mathbf{f}^{(j)}, \mathcal{A}^{(j)} \rangle \right\rangle_{j=1}^{N}$ ($N = 1663$ is the number of sentences), we discriminatively train the frame identification model by maximizing the following log-likelihood:[5]

$$\max_{\boldsymbol{\theta}} \sum_{j=1}^{N} \sum_{i=1}^{m_j} \log \sum_{\ell \in \mathcal{L}_{f_i^{(j)}}} p_{\boldsymbol{\theta}}(f_i^{(j)}, \ell \mid t_i^{(j)}, \mathbf{x}^{(j)}) \quad (3)$$

Note that the training problem is non-convex because of the summed-out prototype latent variable ℓ for each frame. To calculate the objective function, we need to cope with a sum over frames and prototypes for each target (see Eq. 2), often an expensive operation. We locally optimize the function using a distributed implementation of L-BFGS. This is the most expensive model that we train: with 100 CPUs, training takes several hours. (Decoding takes only a few minutes on one CPU for the test set.)

4.4 Results

We evaluate the performance of our frame identification model given gold-standard targets and automatically identified targets (§3); see Table 4.

Given gold-standard targets, our model is able to predict frames for lemmas not seen in training, of which there are 210. The partial-match evaluation gives our model some credit for 190 of these, 4 of which are exactly correct. The hidden variable model, then, is finding related (but rarely exact) frames for unknown target words. The net effect of our conservative target identifier on F_1 is actually positive: the frame identifier is far more precise for targets seen explicitly in training. Together, our target and frame identification outperform the baseline by 4 F_1 points. To compare the frame identification stage in isolation with that of J&N'07, we ran our frame identification model with the targets identified by their system as input. With partial matching, our model achieves a relative improvement of 0.6% F_1 over J&N'07 (though this is not significant).

While our frame identification model thus performs on par with the current state of the art for this task, it improves upon J&N's formulation of the problem because it requires only a single model, learns lexical-semantic features as part of that model rather than requiring a preprocessing step to expand the vocabulary of frame-evoking words, and is probabilistic, which can facilitate global reasoning.

5 Argument Identification

Given a sentence $\mathbf{x} = \langle x_1, \ldots, x_n \rangle$, the set of targets $\mathbf{t} = \langle t_1, \ldots, t_m \rangle$, and a list of evoked frames

[5] We found no benefit on development data from using an L_2 regularizer (zero-mean Gaussian prior).

$\mathbf{f} = \langle f_1, \ldots, f_m \rangle$ corresponding to each target, argument identification is the task of choosing which of each f_i's roles are filled, and by which parts of $\mathbf{x}$. This task is most similar to the problem of semantic role labeling, but uses frame-specific labels that are richer than the PropBank annotations.

5.1 Model

Let $\mathcal{R}_{f_i} = \{r_1, \ldots, r_{|\mathcal{R}_{f_i}|}\}$ denote frame f_i's **roles** (named frame element types) observed in an exemplar sentence and/or our training set. A subset of each frame's roles are marked as **core** roles; these roles are conceptually and/or syntactically necessary for any given use of the frame, though they need not be overt in every sentence involving the frame. These are roughly analogous to the core arguments A0–A5 and AA in PropBank. Non-core roles—analogous to the various AMs in PropBank—loosely correspond to syntactic adjuncts, and carry broadly-applicable information such as the time, place, or purpose of an event. The lexicon imposes some additional structure on roles, including relations to other roles in the same or related frames, and semantic types with respect to a small ontology (marking, for instance, that the entity filling the protagonist role must be sentient for frames of cognition). Fig. 2 illustrates some of the structural elements comprising the frame lexicon by considering the CAUSE_TO_MAKE_NOISE frame.

We identify a set $\mathcal{S}$ of spans that are candidates for filling any role $r \in \mathcal{R}_{f_i}$. In principle, $\mathcal{S}$ could contain any subsequence of $\mathbf{x}$, but in this work we only consider the set of contiguous spans that (a) contain a single word or (b) comprise a valid subtree of a word and all its descendants in the dependency parse produced by the MST parser. This covers 81% of arguments in the development data. The empty span is also included in $\mathcal{S}$, since some roles are not explicitly filled; in the development data, the average number of roles an evoked frame defines is 6.7, but the average number of overt arguments is only 1.7.[6] In training, if a labeled argument is not a valid subtree of the dependency parse, we add its span to $\mathcal{S}$.

Let $\mathcal{A}_i$ denote the mapping of roles in $\mathcal{R}_{f_i}$ to spans in $\mathcal{S}$. Our model makes a prediction for each $\mathcal{A}_i(r_k)$ (for all roles $r_k \in \mathcal{R}_{f_i}$) using:

$$\mathcal{A}_i(r_k) \leftarrow \operatorname{argmax}_{s \in \mathcal{S}} p(s \mid r_k, f_i, t_i, \mathbf{x}) \qquad (4)$$

We use a conditional log-linear model over spans for each role of each evoked frame:

$$p_{\boldsymbol{\psi}}(\mathcal{A}_i(r_k) = s \mid f_i, t_i, \mathbf{x}) = \qquad (5)$$
$$\frac{\exp \boldsymbol{\psi}^\top \mathbf{h}(s, r_k, f_i, t_i, \mathbf{x})}{\sum_{s' \in \mathcal{S}} \exp \boldsymbol{\psi}^\top \mathbf{h}(s', r_k, f_i, t_i, \mathbf{x})}$$

Note that our model chooses the span for each role separately from the other roles and ignores all frames except the frame the role belongs to. Our model departs from the traditional SRL literature by modeling the argument identification problem in a single stage, rather than first classifying token spans as arguments and then labeling them. A constraint implicit in our formulation restricts each role to have at most one overt argument, which is consistent with 96.5% of the role instances in the training data.

Out of the overt argument spans in the training data, 12% are duplicates, having been used by some previous frame in the sentence (supposing some arbitrary ordering of frames). Our role-filling model, unlike a sentence-global argument detection-and-classification approach,[7] permits this sort of argument sharing among frames. The incidence of span *overlap* among frames is much higher; Fig. 1 illustrates a case with a high degree of overlap. Word tokens belong to an average of 1.6 argument spans each, including the quarter of words that do not belong to any argument.

Features for our log-linear model (Eq. 5) depend on the preprocessed sentence $\mathbf{x}$; the target t; a role r of frame f; and a candidate argument span $s \in \mathcal{S}$. Our model includes lexicalized and unlexicalized features considering aspects of the syntactic parse (most notably the dependency path in the parse from the target to the argument); voice; word ordering/overlap/distance of the argument with respect to the target; and POS tags within and around the argument. Many features have a version specific to the frame and role, plus a smoothed version incorporating the role name, but not the frame. These features

[6] In the annotated data, each core role is filled with one of three types of *null instantiations* indicating how the role is conveyed implicitly. E.g., the imperative construction implicitly designates a role as filled by the addressee, and the corresponding filler is thus CNI (constructional null instantiation). In this work we do not distinguish different types of null instantiations.

[7] J&N'07, like us, identify arguments for each target.

are fully enumerated in (Das et al., 2010); instantiating them for our data yields 1,297,857 parameters.

5.2 Training

We train the argument identification model by:

$$\max_{\boldsymbol{\psi}} \sum_{j=1}^{N} \sum_{i=1}^{m_j} \sum_{k=1}^{|\mathcal{R}_{f_i^{(j)}}|} \log p_{\boldsymbol{\psi}}(\mathcal{A}_i^{(j)}(r_k) \mid f_i^{(j)}, t_i^{(j)}, \mathbf{x}^{(j)}) \tag{6}$$

This objective function is concave, and we globally optimize it using stochastic gradient ascent (Bottou, 2004). We train this model until the argument identification F_1 score stops increasing on the development data. Best results on this dataset were obtained with a batch size of 2 and 23 passes through the data.

5.3 Approximate Joint Decoding

Naïve prediction of roles using Eq. 4 may result in overlap among arguments filling different roles of a frame, since the argument identification model fills each role independently of the others. We want to enforce the constraint that two roles of a single frame cannot be filled by overlapping spans. We disallow illegal overlap using a 10000-hypothesis beam search; the algorithm is given in (Das et al., 2010).

5.4 Results

Performance of the argument identification model is presented in Table 5. The table shows how performance varies given different types of perfect input: correct targets, correct frames, and the set of correct spans; correct targets and frames, with the heuristically-constructed set of candidate spans; correct targets only, with model frames; and ultimately, no oracle input (the full frame parsing scenario).

The first four rows of results isolate the argument identification task from the frame identification task. Given gold targets and frames and an oracle set of argument spans, our local model achieves about 87% precision and 75% recall. Beam search decoding to eliminate illegal argument assignments within a frame (§5.3) further improves precision by about 1.6%, with negligible harm to recall. Note that 96.5% recall is possible under the constraint that roles are not multiply-filled (§5.1); there is thus considerable room for improvement with this constraint in place. Joint prediction of each frame's arguments is worth exploring to capture correlations not encoded in our local models or joint decoding scheme.

The 15-point drop in recall when the heuristically-built candidate argument set replaces the set of true argument spans is unsurprising: an estimated 19% of correct arguments are excluded because they are neither single words nor complete subtrees (see §5.1). Qualitatively, the problem of candidate span recall seems to be largely due to syntactic parse errors.[8] Still, the 10-point decrease in precision when using the syntactic parse to determine candidate spans suggests that the model has trouble discriminating between good and bad arguments, and that additional feature engineering or jointly decoding arguments of a sentence's frames may be beneficial in this regard.

The fifth and sixth rows show the effect of automatic frame identification on overall frame parsing performance. There is a 22% decrease in F_1 (18% when partial credit is given for related frames), suggesting that improved frame identification or joint prediction of frames and arguments is likely to have a sizeable impact on overall performance.

The final two rows of the table compare our full model (target, frame, and argument identification) with the baseline, showing significant improvement of more than 4.4 F_1 points for both exact and partial frame matching. As with frame identification, we compared the argument identification stage with that of J&N'07 in isolation, using the automatically identified targets and frames from the latter as input to our model. With partial frame matching, this gave us an F_1 score of 48.1% on the test set—significantly better ($p < 0.05$) than 45.6%, the full parsing result from J&N'07. This indicates that our argument identification model—which uses a single discriminative model with a large number of features for role filling (rather than argument labeling)—is more powerful than the previous state of the art.

6 Related work

Since Gildea and Jurafsky (2002) pioneered statistical semantic role labeling, a great deal of com-

[8] Note that, because of our labels-only evaluation scheme (§2.3), arguments missing a word or containing an extra word receive no credit. In fact, of the frame roles correctly predicted as having an overt span, the correct span was predicted 66% of the time, while 10% of the time the predicted starting and ending boundaries of the span were off by a total of 1 or 2 words.

ARGUMENT IDENTIFICATION					exact frame matching					
	targets	*frames*	*spans*	*decoding*	*P*	*R*	F_1			
Argument identification (oracle spans)	$*$	$*$	$*$	naïve	86.61	75.11	80.45			
	$*$	$*$	$*$	beam §5.3	88.29	74.77	80.97			
Argument identification (full)	$*$	$*$	model §5	naïve	77.43	60.76	68.09	**partial frame matching**		
	$*$	$*$	model §5	beam §5.3	78.71	60.57	68.46	*P*	*R*	F_1
Parsing (oracle targets)	$*$	model §4	model §5	beam §5.3	49.68	42.82	46.00	57.85	49.86	53.56
Parsing (full)	auto §3	model §4	model §5	beam §5.3	**58.08**	**38.76**	**46.49**	**62.76**	**41.89**	**50.24**
Baseline: J&N'07	*auto*	*model*	*model*	*N/A*	*51.59*	*35.44*	*42.01*	*56.01*	*38.48*	*45.62*

Table 5. Argument identification results. $*$ indicates that gold-standard labels were used for a given pipeline stage. For full parsing, bolded scores indicate significant improvements relative to the baseline ($p < 0.05$).

putational work has investigated predicate-argument structures for semantics. Briefly, we highlight some relevant work, particularly research that has made use of FrameNet. (Note that much related research has focused on PropBank (Kingsbury and Palmer, 2002), a set of shallow predicate-argument annotations for *Wall Street Journal* articles from the Penn Treebank (Marcus et al., 1993); a recent issue of *CL* (Màrquez et al., 2008) was devoted to the subject.)

Most work on frame-semantic role labeling has made use of the exemplar sentences in the FrameNet corpus (see §2.1), each of which is annotated for a single frame and its arguments. On the probabilistic modeling front, Gildea and Jurafsky (2002) presented a discriminative model for arguments given the frame; Thompson et al. (2003) used a generative model for both the frame and its arguments; and Fleischman et al. (2003) first used maximum entropy models to find and label arguments given the frame. Shi and Mihalcea (2004) developed a rule-based system to predict frames and their arguments in text, and Erk and Padó (2006) introduced the Shalmaneser tool, which employs Naïve Bayes classifiers to do the same. Other FrameNet SRL systems (Giuglea and Moschitti, 2006, for instance) have used SVMs. Most of this work was done on an older, smaller version of FrameNet.

Recent work on frame-semantic *parsing*—in which sentences may contain multiple frames to be recognized along with their arguments—has used the SemEval'07 data (Baker et al., 2007). The LTH system of Johansson and Nugues (2007), our baseline (§2.4), performed the best in the SemEval'07 task. Matsubayashi et al. (2009) trained a log-linear model on the SemEval'07 data to evaluate argument identification features exploiting various types of taxonomic relations to generalize over roles. A line of work has sought to extend the coverage of FrameNet by exploiting VerbNet, WordNet, and Wikipedia (Shi and Mihalcea, 2005; Giuglea and Moschitti, 2006; Pennacchiotti et al., 2008; Tonelli and Giuliano, 2009), and projecting entries and annotations within and across languages (Boas, 2002; Fung and Chen, 2004; Padó and Lapata, 2005; Fürstenau and Lapata, 2009). Others have applied frame-semantic structures to question answering, paraphrase/entailment recognition, and information extraction (Narayanan and Harabagiu, 2004; Shen and Lapata, 2007; Padó and Erk, 2005; Burchardt, 2006; Moschitti et al., 2003; Surdeanu et al., 2003).

7 Conclusion

We have provided a supervised model for rich frame-semantic parsing, based on a combination of knowledge from FrameNet, two probabilistic models trained on SemEval'07 data, and expedient heuristics. Our system achieves improvements over the state of the art at each stage of processing and collectively, and is amenable to future extension. Our parser is available for download at `http://www.ark.cs.cmu.edu/SEMAFOR`.

Acknowledgments

We thank Collin Baker, Katrin Erk, Richard Johansson, and Nils Reiter for software, data, evaluation scripts, and methodological details. We thank the reviewers, Alan Black, Ric Crabbe, Michael Ellsworth, Rebecca Hwa, Dan Klein, Russell Lee-Goldman, Dan Roth, Josef Ruppenhofer, and members of the ARK group for helpful comments. This work was supported by DARPA grant NBCH-1080004, NSF grant IIS-0836431, and computational resources provided by Yahoo.

References

C. Baker, M. Ellsworth, and K. Erk. 2007. SemEval-2007 Task 19: frame semantic structure extraction. In *Proc. of SemEval*.

H. C. Boas. 2002. Bilingual FrameNet dictionaries for machine translation. In *Proc. of LREC*.

L. Bottou. 2004. Stochastic learning. In *Advanced Lectures on Machine Learning*. Springer-Verlag.

A. Burchardt, K. Erk, and A. Frank. 2005. A WordNet detour to FrameNet. In B. Fisseni, H.-C. Schmitz, B. Schröder, and P. Wagner, editors, *Sprachtechnologie, mobile Kommunikation und linguistische Resourcen*, volume 8. Peter Lang.

A. Burchardt. 2006. Approaching textual entailment with LFG and FrameNet frames. In *Proc. of the Second PASCAL RTE Challenge Workshop*.

D. Das, N. Schneider, D. Chen, and N. A. Smith. 2010. SEMAFOR 1.0: A probabilistic frame-semantic parser. Technical Report CMU-LTI-10-001, Carnegie Mellon University.

K. Erk and S. Padó. 2006. Shalmaneser - a toolchain for shallow semantic parsing. In *Proc. of LREC*.

C. Fellbaum, editor. 1998. *WordNet: an electronic lexical database*. MIT Press, Cambridge, MA.

C. J. Fillmore, C. R. Johnson, and M. R.L. Petruck. 2003. Background to FrameNet. *International Journal of Lexicography*, 16(3).

C. J. Fillmore. 1982. Frame semantics. In *Linguistics in the Morning Calm*, pages 111–137. Hanshin Publishing Co., Seoul, South Korea.

M. Fleischman, N. Kwon, and E. Hovy. 2003. Maximum entropy models for FrameNet classification. In *Proc. of EMNLP*.

P. Fung and B. Chen. 2004. BiFrameNet: bilingual frame semantics resource construction by cross-lingual induction. In *Proc. of COLING*.

H. Fürstenau and M. Lapata. 2009. Semi-supervised semantic role labeling. In *Proc. of EACL*.

D. Gildea and D. Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3).

A.-M. Giuglea and A. Moschitti. 2006. Shallow semantic parsing based on FrameNet, VerbNet and PropBank. In *Proc. of ECAI 2006*.

R. Johansson and P. Nugues. 2007. LTH: semantic structure extraction using nonprojective dependency trees. In *Proc. of SemEval*.

R. Johansson and P. Nugues. 2008. Dependency-based semantic role labeling of PropBank. In *Proc. of EMNLP*.

P. Kingsbury and M. Palmer. 2002. From TreeBank to PropBank. In *Proc. of LREC*.

M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2).

L. Màrquez, X. Carreras, K. C. Litkowski, and S. Stevenson. 2008. Semantic role labeling: an introduction to the special issue. *Computational Linguistics*, 34(2).

Y. Matsubayashi, N. Okazaki, and J. Tsujii. 2009. A comparative study on generalization of semantic roles in FrameNet. In *Proc. of ACL-IJCNLP*.

R. McDonald, K. Crammer, and F. Pereira. 2005. Online large-margin training of dependency parsers. In *Proc. of ACL*.

A. Moschitti, P. Morărescu, and S. M. Harabagiu. 2003. Open-domain information extraction via automatic semantic labeling. In *Proc. of FLAIRS*.

S. Narayanan and S. Harabagiu. 2004. Question answering based on semantic structures. In *Proc. of COLING*.

S. Padó and K. Erk. 2005. To cause or not to cause: cross-lingual semantic matching for paraphrase modelling. In *Proc. of the Cross-Language Knowledge Induction Workshop*.

S. Padó and M. Lapata. 2005. Cross-linguistic projection of role-semantic information. In *Proc. of HLT-EMNLP*.

M. Pennacchiotti, D. De Cao, R. Basili, D. Croce, and M. Roth. 2008. Automatic induction of FrameNet lexical units. In *Proc. of EMNLP*.

A. Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In *Proc. of EMNLP*.

D. Shen and M. Lapata. 2007. Using semantic roles to improve question answering. In *Proc. of EMNLP-CoNLL*.

L. Shi and R. Mihalcea. 2004. An algorithm for open text semantic parsing. In *Proc. of Workshop on Robust Methods in Analysis of Natural Language Data*.

L. Shi and R. Mihalcea. 2005. Putting pieces together: combining FrameNet, VerbNet and WordNet for robust semantic parsing. In *Computational Linguistics and Intelligent Text Processing: Proc. of CICLing 2005*. Springer-Verlag.

M. Surdeanu, S. Harabagiu, J. Williams, and P. Aarseth. 2003. Using predicate-argument structures for information extraction. In *Proc. of ACL*.

C. A. Thompson, R. Levy, and C. D. Manning. 2003. A generative model for semantic role labeling. In *Proc. of ECML*.

S. Tonelli and C. Giuliano. 2009. Wikipedia as frame information repository. In *Proc. of EMNLP*.