Semantic Proto-Role Labeling

Adam Teichert
Johns Hopkins University
teichert@jhu.edu

Adam Poliak
Johns Hopkins University
azpoliak@cs.jhu.edu

Benjamin Van Durme
Johns Hopkins University
vandurme@cs.jhu.edu

Matthew R. Gormley
Carnegie Mellon University
mgormley@cs.cmu.edu

Abstract

The semantic function tags of Bonial, Stowe, and Palmer (2013) and the ordinal, multi-property annotations of Reisinger et al. (2015) draw inspiration from Dowty's semantic proto-role theory. We approach proto-role labeling as a *multi-label* classification problem and establish strong results for the task by adapting a successful model of traditional semantic role labeling. We achieve a proto-role micro-averaged F1 of 81.7 using gold syntax and explore joint and conditional models of proto-roles and categorical roles. In comparing the effect of Bonial, Stowe, and Palmer's tags to PropBank ArgN-style role labels, we are surprised that neither annotations greatly improve proto-role prediction; however, we observe that ArgN models benefit much from observed syntax and from observed or modeled proto-roles while our models of the semantic function tags do not.

1 Introduction

Dowty (1991) argued against the categorical notion of semantic (thematic) roles, suggesting instead a multi-faceted relationship between an argument and a predicate which he termed *proto-roles*. Traditional categories such as AGENT or PATIENT were replaced with prototypical assumptions of underlying semantic properties; e.g. a PROTO-AGENT is likely to be *aware* and *volitional*. This led Reisinger et al. (2015) to construct a dataset supporting the task of semantic proto-role labeling (SPRL): predicting human responses to questions on individual properties. For example, after reading the examples below, consider properties of what was *led*: Was the argument aware of being led? Was it sentient? Was it willing? Did it instigate the leading?

a) The officer *led* **the convict** to the car.
b) California *led* **the nation** in sales.
c) The guide *led* **John** past the danger.

The SPRL task pursued in this paper is a departure from PropBank (Palmer, Gildea, and Kingsbury 2005) semantic role labeling (SRL) which would annotate all of the above examples with the same verb sense (LEAD.01) and argument role (ARG1). The SPRL questions, however, distinguish between these examples without assigning a categorical label. We expect this contrast to provide an opportunity for synergistic joint modeling of SPRL and SRL.

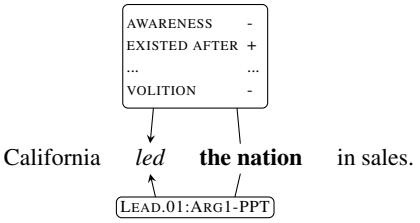


Figure 1: SPRL (top) vs SRL (bottom).

In what follows we:

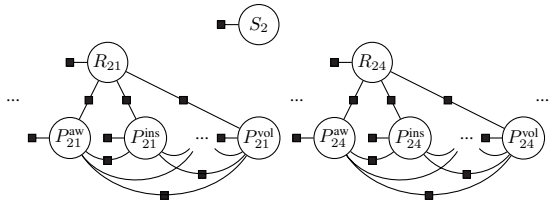
- specify a multi-label classification evaluation for SPRL appropriate for joint labeling of entire input sentences;
- establish a strong SPRL result backed by an SRL model with reasonable performance on a standard dataset;
- evaluate a variety of models with SPRL *and* SRL;
- report SRL results on the new PropBank semantic function tags for Ontonotes 5, contrasting the tagset to PropBank ArgN labels via their impact on our models.

2 Tasks

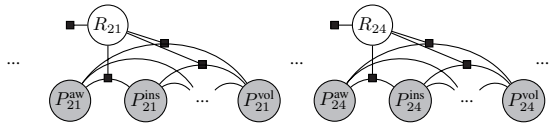
Figure 1 demonstrates the varieties of semantic labeling that we explore in this paper. SRL is traditionally a *multi-class* classification problem where predicate-argument pairs are assigned a label describing the role of the argument in the event. We investigate two label sets for SRL: *ArgN* labels (e.g. Arg0, Arg1) associate the argument to numbered slots for the particular predicate (the numbers tend to hold similar meaning across predicates, but this is not guaranteed); the semantic function tags (*SFT*) of Bonial, Stowe, and Palmer (e.g. PPT, GOL) associate the argument with a coarse-grained role that has meaning across all predicates.

We also investigate proto-role models. We cast SPRL as a *multi-label* classification task where each pair is assigned a *set* of properties (e.g. {awareness, volition}). In the data collected by Reisinger et al., each predicate-argument pair is labeled with two judgments for each SPRL property:

1. Boolean applicability: is the question asked by the property *applicable* to the given pair?
2. A five-way likert response: how likely is it that the property holds in the given context?



(a) SRL+SPRL* with sense variables



(b) SRL|SPRL*: uses pairwise features of observed properties

Figure 2: Factor graphs depicting two models instantiated on the sentence from Fig. 4. S_2 is a sense variable to identify a PropBank frame for the predicate *led*. R_{21} is the SRL variable for the role of **California** with respect to the predicate. P^{aw}_{21},P^{ins}_{21} and P^{vol}_{21} are binary variables representing whether or not **California** has respectively awareness, instigation, and volition in the *led* event ($P^q_{ij} \triangleq P_{ijq}$). R_{24} is the role variable for **nation** in the *led* event.

To formulate the prediction problem as multi-label binary classification, we let the gold label for each predicate-argument pair be the set of properties annotated as "applicable" with likert response of 4 or 5.

3 Models and Features for SPRL and SRL

To establish strong results for SPRL, we take inspiration from the related SRL models of Gormley et al. (2014). We explore several models for three tasks: SRL alone, SPRL alone, and joint prediction of SRL and SPRL. We also optionally include predicate-sense prediction.

Formulation Each model is formalized as a conditional random field or CRF (Lafferty 2001). For each given pred-arg pair (i, j) and property q, we instantiate three types of variables: P_{ijq} is a binary variable with labels $\{+, -\}$ representing that q does or does not hold respectively. R_{ij} is a multi-class variable ranging over SRL role labels. When only the predicate index i is given, we instantiate R_{ij} for all j and allow the role label NIL to indicate that there is no semantic pred-arg relationship. S_i is a multi-class variable ranging over the possible predicate senses. For each model, we select from these variables a task-specific subset:

$$\boldsymbol{Y} = \{Y_k\} \subseteq \{P_{ijq}\} \cup \{R_{ij}\} \cup \{S_i\}.$$

Given the input sentence $\boldsymbol{x}$, the probability of a joint assignment $\boldsymbol{y} = \{y_k\}$ to the variables $\boldsymbol{Y}$ is given by a globally normalized distribution:

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}) \propto \prod_{a \in A} \exp\left(\boldsymbol{w}^T \boldsymbol{f}_a(\boldsymbol{y}_a, \boldsymbol{x})\right),$$

where each $a \in A$ is an index set of variables that some feature looks at jointly, $\boldsymbol{y}_a$ is the corresponding subset of $\boldsymbol{y}$, and $\boldsymbol{w}$ is a vector of parameters. In a factor-graph representation (Frey et al. 1997), A corresponds to the set of factors and defines the independence assumptions.

Models We define five models that vary in two key aspects: the types of variables we include and the structure of the graphical model, given by A.

- SRL includes role variables $\{R_{ij}\}$ with an independent multi-class logistic regression for each—a graphical model with only unary factors.
- SPRL includes property variables $\{P_{ijq}\}$ with an independent binary classifier for each conjunction of predicate-argument pair (i, j) and property q.
- SPRL* has the same variables as SPRL but allows for interactions between pairs of properties. For each pair of properties q and r, there is a factor between P_{ijq} and P_{ijr}.
- SRL+SPRL combines models SRL and SPRL by adding a factor between each SPRL property variable P_{ijq} and its corresponding SRL role variable R_{ij}.
- SRL+SPRL* is our full joint model and includes all factors from models SPRL* and SRL+SPRL. See Figure 2.

The conditional models SRL|SPRL (SRL *given* SPRL) and SPRL|SRL are identical to SRL+SPRL, except that the gold value of each property variable P_{ijq} or role variable R_{ij} is observed respectively—likewise for SPRL*|SRL versus SRL+SPRL*. SRL|SPRL* is identical to SRL|SPRL with the addition of indicator features for each R_{ij} that look at observed *pairs* of SPRL properties.

When evaluating on sense prediction, we also include the sense variables $\{S_i\}$, although they do not share factors with the other variables of the models. We use belief propagation (Pearl 1988; Kschischang, Frey, and Loeliger 2001) for inference. For the models with cycles (SPRL*, SRL+SPRL*), we run loopy belief propagation (Pearl 1988; Murphy, Weiss, and Jordan 1999) with a maximum of five iterations. Our implementation uses the Pacaya library[1].

Features As is typical in CRFs, we define each of our features on a factor a as a conjunction of an indicator $\mathbb{1}$ for a fixed variable assignment $\tilde{\boldsymbol{y}}_a$ with some *observation-feature* function g_{ak} of the input sentence:

$$f_{a,k,\tilde{\boldsymbol{y}}_a}(\boldsymbol{y}_a, \boldsymbol{x}) = \mathbb{1}(\boldsymbol{y}_a = \tilde{\boldsymbol{y}}_a) g_{ak}(\boldsymbol{x}).$$

We include over one hundred observation-features motivated by prior work in dependency-based SRL (Björkelund, Hafdell, and Nugues 2009; Zhao et al. 2009; Lluís, Carreras, and Màrquez 2013). The features use the sentence's words, lemmas, Brown clusters (Brown et al. 1992)[2], part-of-speech tags, and syntactic dependency parse. When present, inter-property and SPRL-SRL factors only include a bias parameter for each configuration. We employ the feature-hashing trick (Ganchev and Dredze 2008; Weinberger et al. 2009) to restrict the number of model parameters.

[1] https://github.com/mgormley/pacaya

[2] We use https://github.com/percyliang/brown-cluster to create 1000 clusters of wikitxt from the polyglot project (Al-Rfou, Perozzi, and Skiena 2013). Our features look at the full id and length-five prefixes.

	annotated sentences	pred-arg instances	# label types ArgN	SFT
CoNLL09	43,012	430,850	53	-
OntoFull	35,497	266,298	31	26
OntoMed	24,755	185,878	31	26
OntoSmall	4,912	36,618	27	24
PropSmall	4,912	9,738	20	16

Table 1: Dataset sizes

Prior work has explored joint syntactic and semantic dependency parsers to understand the interaction between the two linguistic strata (Johansson 2009; Gesmundo et al. 2009; Naradowsky, Riedel, and Smith 2012; Lluís, Carreras, and Màrquez 2013; Gormley et al. 2014). Here, by contrast, we are interested in the relation between different semantic annotation schemes. Nonetheless, our joint model is similar in both form and features.[3]

4 Experiments

Datasets Our experiments use several datasets. PropBank adds semantic role labels to the syntactic annotations available on the Wall Street Journal (WSJ) portion of the Penn Treebank (Marcus, Marcinkiewicz, and Santorini 1993). Each predicate instance in the corpus is labeled with a *verb sense* (a.k.a. *roleset*) which has a corresponding *frame*. See Figure 3 for the frame corresponding to LEAD.01 from our example. Each frame describes the *slots* that can be filled by the predicate's arguments. Arguments of each predicate instance are identified as such and labeled so as to identify which slot it fills. For example, California fills the ARG1 slot in Figure 1. ***PropSmall*** contains the subset of PropBank predicate-argument pairs as filtered and further annotated by Reisinger et al. (2015). This is our only dataset containing SPRL annotations.

In PropBank, the role labels (e.g. Arg1, Arg2) are not necessarily consistent in meaning across rolesets and must be disambiguated by the frame. However, Ontonotes 5 (Weischedel et al. 2013; Bonial, Stowe, and Palmer 2013) — a more recent extension of PropBank[4] — additionally annotates each slot with one of a small number of labels called *propbank semantic function tags* (SFT) whose meanings are not roleset specific. These are shown after the hyphen in the example of Figure 3. Having roleset-independent tags justifies sharing statistical strength of observations across all training examples. The Ontonotes 5 dataset includes most (but not all) of the Penn Treebank WSJ sentences as well as data from other genres. Our experiments on Ontonotes

Roleset: *Lead.01*
Name: *directed motion, be ahead of*
Arg0-**PAG**: leader
Arg1-**PPT**: in the lead of
Arg2-**EXT**: extent
Arg4-**DIR**: start point
Arg5-**GOL**: end point
Examples:
cause to go: John led the unhappy ...
go before: California led the nation ...
...

Figure 3: Example of information available in PropBank framesets (v3.1) for Lead.01.

5 are restricted to the WSJ subset. ***OntoFull*** is composed of all overt predicate-argument pairs in the WSJ portion of Ontonotes 5. It includes SFT annotations in addition to ArgN SRL labels. ***OntoMed*** and ***OntoSmall*** include the pairs from random subsets of the sentences in OntoFull. Figure 1 compares the sizes of our datasets.

The PropBank, Ontonotes, and SPRL datasets were originally annotated relative to constituency parses. We automatically map gold constituency parses to universal Stanford dependencies (de Marneffe et al. 2014) and gold part-of-speech tags to the universal part-of-speech tagset (Petrov, Das, and McDonald 2012). [5]

CoNLL09 is the English SRL data from the CoNLL-2009 shared task (Hajič et al. 2009; Surdeanu et al. 2008) and includes verbal and nominal predicates from PropBank (Palmer, Gildea, and Kingsbury 2005) and NomBank (Meyers et al. 2004) respectively. The English data from the CoNLL-2009 shared task (Hajič et al. 2009; Surdeanu et al. 2008) included head-based semantic role labeling and sense prediction. We use the CoNLL-2009 data to validate the performance of our SRL model.

Training We train our models using stochastic gradient descent (SGD) with the AdaGrad adaptive learning rate and a composite mirror descent objective with ℓ_2 regularization following Duchi, Hazan, and Singer (2011). We used the train data to define the SGD objective and to (optionally) adjust the AdaGrad η parameter during learning (Bottou 2012). We used our evaluation objective (e.g. Labeled SPRL F1) on the dev data for early stopping.[6] Wherever we report aggregated F1 over all properties, it is micro-averaged F1. We used random search for hyper-parameter optimization (Bergstra and Bengio 2012), sampling thirty random configurations.[7] For each model scenario, we trained under

[3]The model of Naradowsky, Riedel, and Smith looks especially similar to ours for SPRL (i.e. they include a collection of binary variables for each pred-arg pair); however, theirs is a *multi-class* model using hard factors to enforce mutual exclusion of the labels and is more akin to our SRL model. Such constraints are inappropriate for multi-label SPRL.

[4]We used the release-candidate version of the frames: https://github.com/propbank/propbank-frames/tree/release-candidate

[5]We use PyStanfordDependencies: https://github.com/dmcc. As the gold head for PropBank and OntoNotes predicates and arguments, we select the left-most token whose parent in the converted gold dep-parse is not in the set of dominated tokens.

[6]Our joint models were each trained in view of optimizing only one objective at a time. That is, the models in Table 4 were trained using labeled SRL accuracy as the evaluation objective while the models in Table 5 used SPRL Property F1.

[7]For each random configuration, hyper-parameters were independetly selected from the following ranges: adaGradEta [5e-4,

	SRL +sense +arg-id Labeled F1	
0	Naradowsky, Riedel, and Smith (2012)	78.55
1	Gormley et al. (2014)	86.54
2	our SRL model	**87.40**
	SRL +sense Accuracy	
3	our SRL model	90.78
	SRL Accuracy	
4	our SRL model	91.48

Table 2: English CoNNL 2009 SRL given gold syntax. Lines 0-1 report published results evaluating labeled role and sense F1 with predicate heads pre-identified; line 2 is our model for the same setting, the line 3 model has predicate-argument pairs pre-identified (so F1=Accuracy), and line 4 drops sense disambiguation from the evaluation.

		OntoFull		**PropSmall**	
	train	ArgN	SFT	ArgN	SFT
0	OntoFull	88.3	87.5	86.1	82.9
1	OntoMed	87.2	86.5	85.8	82.1
2	OntoSmall	82.3	81.4	82.0	77.0
3	PropSmall	-	-	87.0	79.1

Table 3: SRL accuracy given gold syntax and pre-identified predicate-argument pairs under various train/test conditions. Rows correspond to the dataset from which the train data was used. Columns identify the labelset and the data from which the test and dev sets were used.

all hyper-parameter configurations, selected the model with the best dev performance, and evaluated on held out data.

With the exception of CoNLL09, we split the datasets on WSJ section boundaries as follows: train (0-18), dev (19-21), test (22-24). To compensate for the smaller size of the PropSmall dataset which was filtered and sampled from PropBank by Reisinger et al., our split reserves a larger proportion of the data for development and test than does CoNLL09.

SRL Table 2 shows that our SRL model performs well compared to published work on the English CoNLL-2009 task using gold dependencies and part-of-speech tags. It also shows the baseline performance on the SRL task we use in the remainder of the paper (i.e. gold predicate-argument pairs are pre-identified and predicate sense is not evaluated). We include the two baselines from the literature of which we are aware that use gold syntax for English CoNLL-2009.

Table 3 provides insights into the PropSmall SRL data and contrasts the ArgN and SFT labelsets. Unsurprisingly, regardless of the labelset, our SRL models perform worse when fewer training examples are available. When train and test are both from a random sample of Ontonotes (i.e. the OntoFull columns) the degradation as a function of training size is roughly independent of the tagset. However, training on the random subsets and testing on PropSmall hurts SFT prediction (> 4.4 decrease) more than ArgN (< 2.3 decrease). Rows 2 and 3 show a large contrast between ArgN and SFT prediction on the two datasets.

1.0], L2Lambda [1e-10, 10], featCountCutoff {1,2,3,4}, sgdAutoSelectLr {True, False}. Continuous parameters were sampled on a log scale and then rounded to 2 significant digits.

		syntax=gold		**syntax=none**	
	setting	ArgN	SFT	ArgN	SFT
0	SRL	87.0	79.1	82.7	79.1
1	SRL|SPRL	87.7	80.2	83.2	80.4
2	SRL|SPRL*	86.8	80.5	84.5	79.4
3	SRL+SPRL	86.5	80.7	84.4	78.4
4	SRL+SPRL*	86.3	79.8	83.8	78.1

Table 4: Accuracy of SRL argument labeling in isolation, given SPRL, or modeled jointly with SPRL; * indicates second-order SPRL features.

		syntax=gold		**syntax=none**	
	setting	ArgN	SFT	ArgN	SFT
0	SPRL	80.9		80.7	
1	SPRL*	81.7		80.8	
2	SPRL|SRL	81.5	81.4	82.0	81.4
3	SPRL*|SRL	81.8	80.8	81.7	81.8
4	SRL+SPRL	81.2	81.1	81.0	80.9
5	SRL+SPRL*	81.3	81.3	81.2	81.1

Table 5: Multi-label F1 of SPRL in isolation, given SRL, or modeled jointly with SRL.

SRL using SPRL Table 4 shows the SRL results on test data from models that incorporate varying amounts of SPRL information. The SRL model uses no SPRL annotations, SRL|SPRL and SRL|SPRL* use gold annotations at test time, while SRL+SPRL and SRL+SPRL* only use SPRL annotations at training time. Intuitively, SPRL set-valued labels provide refinements of the coarser-grained SRL labels. Comparing rows 0 and 1, we see that in all cases, features of observed gold SPRL annotations allow us to learn better models. Our results are mixed for adding higher-order features and for jointly modeling SRL with SPRL. Comparing row 0 to rows 3-4, we see inferred SPRL helping SFT labeling when gold syntax is available and helping ArgN labeling when syntax is not available.[8]

SPRL with SRL Table 5 shows results for SPRL evaluated as a retrieval task with F1. The results of these models are much more invariant to the availability of our syntactic features than were the SRL results of Table 4. The models with second-order property factors in row 1 improve over

[8] In dev results (not shown here), the row 1 models excel those of row 0 by even larger margins than on test while rows 3 and 4 actually perform worse than those of row 0.

Baseline	F1	Prec	Rec
property majority	59.1	70.4	50.9
max type-level F1	62.9	48.9	88.3
Reisinger et al. (2015)	71.0	67.9	74.4

Reisinger et al. (2015)					possible	
By Property	**CF**	**F1**	**Prec**	**Rec**	**train**	**dev**
instigation	+	76.7	63.3	97.3	2811	376
volition	+	69.8	56.4	91.6	2728	350
awareness	+	68.8	57.4	85.7	3021	390
sentient	0	42.0	54.5	34.1	1856	244
physically existed	0	50.0	44.4	57.1	2663	362
existed before	+	79.5	67.9	95.9	4978	699
existed during	+	93.1	89.2	97.4	6566	879
existed after	+	82.3	71.1	97.7	5358	729
created	-	0.0	100.0	0.0	549	73
destroyed	-	17.1	33.3	11.5	230	40
changed	0	54.0	61.4	48.2	2735	400
changed state	0	54.6	61.3	49.2	2705	396
changed possession	-	0.0	100.0	0.0	473	74
changed location	-	6.6	66.7	3.4	575	55
stationary	-	13.3	40.0	8.0	285	53
location	-	0.0	100.0	0.0	621	82
physical contact	-	21.5	48.5	13.8	1138	150
manipulated	+	72.1	80.9	65.1	4048	606

Table 6: Aggregate multi-label SPRL results and breakdown by property the Reisinger et al. model. **CF** is to aid visualization: + for F1 > 66.7, - for F1 < 33.3 and 0 otherwise. The rightmost columns report the number of positive instances in the gold train and dev sections of PropSmall.

those without in row 0. Conditioning on gold SRL or jointly modeling SRL and SPRL generally helps except in some cases where the second-order property factors are present.

SPRL Baselines To the best of our knowledge, the work of Reisinger et al. (2015) contains the only prior SPRL result and, according to personal correspondence with some of the authors, their predictive models were not a primary goal of that work. A key contribution of this paper is that we refine the evaluation and propose a model that substantially outperforms the previously evaluated models. We have modified the dataset split so as to be amenable to joint modeling at the sentence (or even the section) level which makes the prediction results released with the dataset (Reisinger et al. 2015) not directly comparable to ours. Therefore, Table 6 replicates the approach of Reisinger et al. (2015) using our evaluation and includes two other SPRL baselines (compare to Tables 5 and 7; e.g. 71.0 versus 81.7 F1 from our model). Our re-implementation of the "Full" method in Reisinger et al. (2015) uses LibLinear (Fan et al. 2008) to fit a linear model with a property-specific bias, a feature encoding the distance and direction from the predicate to the argument and an embedding of the predicate. We tuned a property-specific regularization coefficient on dev aggregate F1.[9] The table also includes two additional aggregate baselines that assign labels at the *type* level (i.e. each property is either predicted as always present or absent). The first assigns the majority label for each property according to the train+dev data. The second assigns a positive label to the k most frequent properties and then optimizes k for F1 on train+dev ($k = 10$ being the best). After using the train and dev data to determine which properties to always predict as positive, we evaluate those predictions on the test data.

Property	CF	F1	Prec	Rec	possible train	dev
instigation	+	85.6	83.1	88.3	2811	376
volition	+	86.4	84.3	88.5	2728	350
awareness	+	87.3	85.7	88.9	3021	390
sentient	+	85.6	88.1	83.2	1856	244
physically existed	+	76.4	79.3	73.8	2663	362
existed before	+	84.8	84.1	85.6	4978	699
existed during	+	95.1	93.0	97.2	6566	879
existed after	+	87.5	84.7	90.5	5358	729
created	0	44.4	64.9	33.8	549	73
destroyed	-	0.0	0.0	0.0	230	40
changed	+	67.8	67.5	68.2	2735	400
changed state	0	66.1	67.8	64.4	2705	396
changed possession	0	38.8	87.0	25.0	473	74
changed location	0	35.6	86.7	22.4	575	55
stationary	-	21.4	100.0	12.0	285	53
location	-	18.5	58.8	11.0	621	82
physical contact	0	40.7	62.5	30.2	1138	150
manipulated	+	86.0	85.4	86.6	4048	606
total		81.7	83.1	80.3		

Table 7: Breakdown of SPRL* results on held out test data with gold syntax. Compare to baselines in Table 6.

SPRL Breakdown By Property We now take a closer look at results from our best SPRL model, SPRL* with gold syntax (81.7 held-out F1). Table 7 gives a breakdown of results by property (compare to baselines in Table 6 and aggregate results in Table 5). As with the Reisinger baseline, our best performance (95.1) is for EXISTED DURING while we get less than 30.0 F1 for DESTROYED, STATIONARY and LOCATION. Clearly, we struggle most with predicting the presence of infrequent properties. This is not surprising since our micro-averaged F1 metric on which we tuned hyperparameters encourages us to focus on the categories with the most examples.

SPRL Examples Figure 4 shows a variety of cherry-picked outputs from the model on dev examples. In (a) it is unclear whether **two Boston sales representatives** should actually be considered the location of the event. In (b) our model infers that the **shops** did not exist until they were

[9] In contrast, tuning a single regularization coefficient (as we did for our other models) resulted in worse held-out F1 which is made even worse if property-specific bias features are included.

a In August , soon after ... replaced its president , **two Boston sales representatives** *sent* customers a letter saying ...

b Last year , the Irish airport authority , in a joint venture with Aeroflot , *opened* **four hard-currency duty-free shops** ...

c Enormous ice sheets *retreated* from the face of North America , northern Europe and Asia .

d The notice also *grants* **relief** for certain estate-tax returns .

e ... **a reporter for the Reuters newswire** *miscalculated* the industrial average 's drop as a 4 % decline ...

f She and her husband *started* a small printing business and need the car for work as well as for weekend jaunts .

g In 1979 , **the pair** *split* the company in half , with Walter and his son , Sam , agreeing to operate under ...

h Mr. Paul denies phoning and *gloating* .

Property	a	b	c	d	e	f	g	h
instigation	+				+	+	+	+
volition	+				+	+	+	+
awareness	+				+	+	+	+
sentient	+				+	+		+
physically existed	+	+			+	+		+
existed before	+		+		+	+	+	+
existed during	+	+	+	+	+	+	+	+
existed after	+	+	+	+	+	+	+	+
created		+						
destroyed								
changed		+	+	+		+	+	+
changed state		+	+			+	+	
changed possession			+					
changed location						+		
stationary								
location								
physical contact	+	+				+		
manipulated		+	+	+				

Figure 4: SPRL predictions from SPRL* with gold syntax; highlighted cells reflect disagreement with annotator.

opened. Our output for (d) oddly misses that **relief** was CREATED despite correctly identifying that it did not EXIST BEFORE and did EXIST AFTER. In (g), it is surprising that the model correctly predicts VOLITION and AWARENESS but misses SENTIENT. This might be due to incorrect signal from also missing PHYSICALLY EXISTED. Conversely, in the *miscalculated* predicate of example (e) the annotations identify a **reporter** that is SENTIENT and has VOLITION but not AWARENESS, while the model infers that AWARENESS indeed holds. Example (h) deals with a tricky, dubious event.

5 Related Work

The evaluation of Reisinger et al. accompanying the SPRL data release is the most closely related to our work. However, our experiments address several concerns with their setup. We split the data on section boundaries rather than randomly selecting predicate-argument pairs. We incorporate features from the SRL literature and allow properties to be predicted jointly, whereas their setup used a deliberately simple set of features and predicted properties independently. Our treatment of SPRL as multi-label classification also leads to a different evaluation metric. Table 6 shows the baseline for the new data splits and evaluation metric. Several authors have considered trade-offs in annotator effort and data-sparsity in arising in traditional SRL annotations (Loper, Yi, and Palmer 2007; Yi, Loper, and Palmer 2007; Zapirain, Agirre, and Màrquez 2008).

6 Conclusions and Future Work

We established the best reported results for SPRL under a simple multi-label classification paradigm when predicate-argument pairs have already been identified. We sought improvements to our SPRL model by including pairwise proto-role factors and factors that join categorical role variables with proto-role variables. We also investigated the contrast between ArgN and semantic function tags as the underlying theory for categorical role labeling and we looked at the importance of dependency parse information to the model.

Suprisingly, we find that obsevered syntax and semantic roles give little boost to SPRL F1 (at most, 1.3 absolute) and that SFT SRL prediction also gains relatively little from using SPRL or syntax. These negative results deserve further investigation. We believe that future work into improved joint models should show stronger interactions between SRL and SPRL. In contrast, our best ArgN SRL model on the same predicate-argument instances makes large gains of 4.5 absolute F1 over the syntax-free analog and 0.7 over the analog without SPRL. Furthermore, when syntax is not available, ArgN SRL benefits from SPRL annotations available only at training time, improving by 1.7 absolute F1.

In future work, we wish to better leverage the ordinal nature of the collected responses and handle the SPR2.x data (White et al. 2016) that includes multiple, overlapping annotators.

Acknowledgments

We thank Drew Reisinger for assistance with the SPRL data and Tim O'Gorman for his help with PropBank and for providing function tag labels for the SPRL annotated sentences that are not in Ontonotes 5. We also thank colleagues and anonymous reviewers for helpful discussion and feedback. This research was supported by the Human Language Technology Center of Excellence (HLTCOE), and DARPA LORELEI. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes. The views and conclusions contained in this publication are those of the authors and should not be interpreted as representing official policies or endorsements of DARPA or the U.S. Government.

References

Al-Rfou, R.; Perozzi, B.; and Skiena, S. 2013. Polyglot: Distributed word representations for multilingual nlp. In *Proceedings of CoNLL*, 183–192.

Bergstra, J., and Bengio, Y. 2012. Random search for hyper-parameter optimization. *Journal of Machine Learning Research* 13(1):281–305.

Björkelund, A.; Hafdell, L.; and Nugues, P. 2009. Multilingual semantic role labeling. In *Proceedings of CoNLL*, 43–48.

Bonial, C.; Stowe, K.; and Palmer, M. 2013. Renewing and revising semlink. In *The GenLex Workshop on Linked Data in Linguistics*.

Bottou, L. 2012. Stochastic gradient descent tricks. In *Neural Networks: Tricks of the Trade*. Springer. 421–436.

Brown, P. F.; Desouza, P. V.; Mercer, R. L.; Pietra, V. J. D.; and Lai, J. C. 1992. Class-based n-gram models of natural language. *Computational Linguistics* 18(4):467–479.

de Marneffe, M.-C.; Dozat, T.; Silveira, N.; Haverinen, K.; Ginter, F.; Nivre, J.; and Manning, C. D. 2014. Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of LREC*, 4585–4592.

Dowty, D. 1991. Thematic proto-roles and argument selection. *Language* 67(3):547–619.

Duchi, J.; Hazan, E.; and Singer, Y. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research* 12:2121–2159.

Fan, R.-E.; Chang, K.-W.; Hsieh, C.-J.; Wang, X.-R.; and Lin, C.-J. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research* 9:1871–1874.

Frey, B. J.; Kschischang, F. R.; Loeliger, H.-A.; and Wiberg, N. 1997. Factor graphs and algorithms. In *Proceedings of the Annual Allerton Conference on Communication Control and Computing*, volume 35.

Ganchev, K., and Dredze, M. 2008. Small statistical models by random feature mixing. In *Proceedings of the ACL08 HLT Workshop on Mobile Language Processing*, 19–20.

Gesmundo, A.; Henderson, J.; Merlo, P.; and Titov, I. 2009. A latent variable model of synchronous syntactic-semantic parsing for multiple languages. In *Proceedings of CoNLL*, 37–42.

Gormley, M. R.; Mitchell, M.; Van Durme, B.; and Dredze, M. 2014. Low-resource semantic role labeling. In *Proceedings of ACL*.

Hajič, J.; Ciaramita, M.; Johansson, R.; Kawahara, D.; Martí, M. A.; Màrquez, L.; Meyers, A.; Nivre, J.; Padó, S.; Štěpánek, J.; et al. 2009. The conll-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of CoNLL*, 1–18. Association for Computational Linguistics.

Johansson, R. 2009. Statistical bistratal dependency parsing. In *Proceedings of EMNLP*, 561–569.

Kschischang, F. R.; Frey, B. J.; and Loeliger, H.-A. 2001. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory* 47(2).

Lafferty, J. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML*. Morgan Kaufmann.

Lluís, X.; Carreras, X.; and Màrquez, L. 2013. Joint arc-factored parsing of syntactic and semantic dependencies. *TACL* 1:219–230.

Loper, E.; Yi, S.-T.; and Palmer, M. 2007. Combining lexical resources: mapping between propbank and verbnet. In *Proceedings of the 7th International Workshop on Computational Linguistics*.

Marcus, M.; Marcinkiewicz, M.; and Santorini, B. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19(2):330.

Meyers, A.; Reeves, R.; Macleod, C.; Szekely, R.; Zielinska, V.; Young, B.; and Grishman, R. 2004. The nombank project: An interim report. In *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, 24–31.

Murphy, K. P.; Weiss, Y.; and Jordan, M. I. 1999. Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of UAI*, 467–475. Morgan Kaufmann Publishers Inc.

Naradowsky, J.; Riedel, S.; and Smith, D. A. 2012. Improving nlp through marginalization of hidden syntactic structure. In *Proceedings of EMNLP-CoNLL*, 810–820. Association for Computational Linguistics.

Palmer, M.; Gildea, D.; and Kingsbury, P. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics* 31(1):71–106.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Petrov, S.; Das, D.; and McDonald, R. 2012. A universal part-of-speech tagset. In *Proceedings of LREC*.

Reisinger, D.; Rudinger, R.; Ferraro, F.; Harman, C.; Rawlins, K.; and Durme, B. V. 2015. Semantic proto-roles. *TACL* 3:475–488.

Surdeanu, M.; Johansson, R.; Meyers, A.; Màrquez, L.; and Nivre, J. 2008. The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of CoNLL*.

Weinberger, K.; Dasgupta, A.; Langford, J.; Smola, A.; and Attenberg, J. 2009. Feature hashing for large scale multitask learning. In *Proceedings of ICML*, 1113–1120.

Weischedel, R.; Palmer, M.; Marcus, M.; Hovy, E.; Pradhan, S.; Ramshaw, L.; Xue, N.; Taylor, A.; Kaufman, J.; Franchini, M.; El-Bachouti, M.; Belvin, R.; and Houston, A. 2013. Ontonotes release 5.0.

White, S. A.; Reisinger, D.; Sakaguchi, K.; Vieira, T.; Zhang, S.; Rudinger, R.; Rawlins, K.; and Van Durme, B. 2016. Universal decompositional semantics on universal dependencies. In *Proceedings of EMNLP*, 1713–1723.

Yi, S.-T.; Loper, E.; and Palmer, M. 2007. Can semantic roles generalize across genres? In *Proceedings of HLT-NAACL*, 548–555.

Zapirain, B.; Agirre, E.; and Màrquez, L. 2008. Robustness and generalization of role sets: PropBank vs. VerbNet. In *Proceedings of ACL-08: HLT*, 550–558.

Zhao, H.; Chen, W.; Kity, C.; and Zhou, G. 2009. Multilingual dependency learning: A huge feature engineering method to semantic dependency parsing. In *Proceedings of CoNNL*, 55–60.