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.

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?

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.
variables: \( P \)

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. \( P^\text{aw}, P^\text{ins}, \) and \( P^\text{vol} \) are binary variables representing whether or not California has respectively awareness, instigation, and volition in the led event (\( P^\text{aw}_i \approx P^\text{vol}_i \)). \( R_{24} \) 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. \( P_{ij} \) 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:

\[ Y = \{Y_k\} \subseteq \{P_{ijq}\} \cup \{R_{ij}\} \cup \{S_i\}. \]

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

\[ p(y|x, w) \propto \prod_{a \in A} \exp \left( w^T f_a(y_a, x) \right), \]

where each \(a \in A\) is an index set of variables that some feature looks at jointly, \(y_a\) is the corresponding subset of \(y\), and \(w\) 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 SPRL|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.

Features As is typical in CRFs, we define each of our features on a factor \(a\) as a conjunction of an indicator \(I\) for a fixed variable assignment \(y_a\) with some observation-feature function \(g_{a,k}\) of the input sentence:

\[ f_{a,k}(y_a, x) = I(y_a = \tilde{y}_a)g_{a,k}(x). \]

We include over one hundred observation-features motivated by prior work in dependency-based SRL (Björkelund, Hafðell, and Nuges 2009; Zhao et al. 2009; Lluíà, Carreras, and Márquez 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.

\(^*\)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.
Reisinger et al. (2015). This is our only dataset containing predicate-argument pairs as filtered and further annotated by slot in Figure 1. PropBank is appropriate for multi-label SPRL.

Naradowsky, Riedel, and Smith 2012; Lluis, Carreras, and Márquez 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.

### 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 — 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 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, Dus, and McDonald 2012). 5

### Table 1: Dataset sizes

<table>
<thead>
<tr>
<th>Dataset</th>
<th>annotated sentences</th>
<th>pred-arg instances</th>
<th># label types</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoNLL09</td>
<td>43,012</td>
<td>430,850</td>
<td>ArgN SFT</td>
</tr>
<tr>
<td>OntoFull</td>
<td>35,497</td>
<td>266,298</td>
<td>31 26</td>
</tr>
<tr>
<td>OntoMed</td>
<td>24,755</td>
<td>185,878</td>
<td>31 26</td>
</tr>
<tr>
<td>OntoSmall</td>
<td>4,912</td>
<td>36,618</td>
<td>27 24</td>
</tr>
<tr>
<td>PropSmall</td>
<td>4,912</td>
<td>9,738</td>
<td>20 16</td>
</tr>
</tbody>
</table>

- **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.*

#### 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 ...

- Wherever we report aggregated F1 over all properties, it is micro-averaged F1. For each model scenario, we trained under independent selected from the following ranges: adaGradEta [5e-4,
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.

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.

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

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.

Table 4: Accuracy of SRL argument labeling in isolation, given SPRL, or modeled jointly with SPRL; * indicates second-order SPRL features.
Evaluation and includes two other SPRL baselines (compare replicates the approach of Reisinger et al. (2015) using our al. 2015) not directly comparable to ours. Therefore, Table 6 the prediction results released with the dataset (Reisinger et the sentence (or even the section) level which makes modified the dataset split so as to be amenable to joint mod-
tially outperforms the previously evaluated models. We have refine the evaluation and propose a model that substan-
goal of that work. A key contribution of this paper is that of Reisinger et al. (2015) contains the only prior SPRL re-
result and, according to personal correspondence with some 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. 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 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.

**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

<table>
<thead>
<tr>
<th>Baseline</th>
<th>F1</th>
<th>Prec</th>
<th>Rec</th>
</tr>
</thead>
<tbody>
<tr>
<td>property majority</td>
<td>59.1</td>
<td>70.4</td>
<td>50.9</td>
</tr>
<tr>
<td>max type-level F1</td>
<td>62.9</td>
<td>48.9</td>
<td>88.3</td>
</tr>
<tr>
<td>Reisinger et al. (2015)</td>
<td>71.0</td>
<td>67.9</td>
<td>74.4</td>
</tr>
</tbody>
</table>

**Reisinger et al. (2015) By Property**

<table>
<thead>
<tr>
<th>Property</th>
<th>CF</th>
<th>F1</th>
<th>Prec</th>
<th>Rec</th>
<th>train</th>
<th>dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>instigation</td>
<td>+</td>
<td>76.7</td>
<td>63.3</td>
<td>97.3</td>
<td>2811</td>
<td>376</td>
</tr>
<tr>
<td>volition</td>
<td>+</td>
<td>69.8</td>
<td>56.4</td>
<td>91.6</td>
<td>2728</td>
<td>350</td>
</tr>
<tr>
<td>awareness</td>
<td>+</td>
<td>68.8</td>
<td>57.4</td>
<td>85.7</td>
<td>3021</td>
<td>390</td>
</tr>
<tr>
<td>sentient</td>
<td>0</td>
<td>42.0</td>
<td>54.5</td>
<td>34.1</td>
<td>1856</td>
<td>244</td>
</tr>
<tr>
<td>physically existed</td>
<td>0</td>
<td>50.0</td>
<td>44.4</td>
<td>57.1</td>
<td>2663</td>
<td>362</td>
</tr>
<tr>
<td>existed before</td>
<td>+</td>
<td>79.5</td>
<td>67.9</td>
<td>95.9</td>
<td>4978</td>
<td>699</td>
</tr>
<tr>
<td>existed during</td>
<td>+</td>
<td>93.1</td>
<td>89.2</td>
<td>97.4</td>
<td>6566</td>
<td>879</td>
</tr>
<tr>
<td>existed after</td>
<td>+</td>
<td>82.3</td>
<td>71.1</td>
<td>97.7</td>
<td>5358</td>
<td>729</td>
</tr>
<tr>
<td>created</td>
<td>-</td>
<td>0.0</td>
<td>100.0</td>
<td>0.0</td>
<td>549</td>
<td>73</td>
</tr>
<tr>
<td>destroyed</td>
<td>-</td>
<td>17.1</td>
<td>33.3</td>
<td>11.5</td>
<td>230</td>
<td>40</td>
</tr>
<tr>
<td>changed</td>
<td>0</td>
<td>54.0</td>
<td>61.4</td>
<td>48.2</td>
<td>2735</td>
<td>400</td>
</tr>
<tr>
<td>changed state</td>
<td>0</td>
<td>54.6</td>
<td>61.3</td>
<td>49.2</td>
<td>2705</td>
<td>396</td>
</tr>
<tr>
<td>changed possession</td>
<td>0</td>
<td>0.0</td>
<td>100.0</td>
<td>0.0</td>
<td>473</td>
<td>74</td>
</tr>
<tr>
<td>changed location</td>
<td>-</td>
<td>6.6</td>
<td>66.7</td>
<td>3.4</td>
<td>575</td>
<td>77</td>
</tr>
<tr>
<td>stationary</td>
<td>-</td>
<td>13.3</td>
<td>40.0</td>
<td>8.0</td>
<td>285</td>
<td>53</td>
</tr>
<tr>
<td>location</td>
<td>-</td>
<td>0.0</td>
<td>100.0</td>
<td>0.0</td>
<td>621</td>
<td>82</td>
</tr>
<tr>
<td>physical contact</td>
<td>-</td>
<td>21.5</td>
<td>48.5</td>
<td>13.8</td>
<td>1138</td>
<td>150</td>
</tr>
<tr>
<td>manipulated</td>
<td>+</td>
<td>72.1</td>
<td>80.9</td>
<td>65.1</td>
<td>4048</td>
<td>606</td>
</tr>
</tbody>
</table>

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.

Table 7: Breakdown of SPRL* results on held out test data with gold syntax. Compare to baselines in Table 6.
in August, soon after replacement its president, two Boston sales representatives sent customers a letter saying...

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

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

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

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

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

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

Mr. Paul denies phoning and gloating...

<table>
<thead>
<tr>
<th>Property</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
<th>g</th>
<th>h</th>
</tr>
</thead>
<tbody>
<tr>
<td>instigation</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>volition</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>awareness</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>sentient</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>physically existed</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>existed before</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>existed after</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>created</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>destroyed</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>changed</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>changed state</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>changed possession</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>stationary</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>location</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>physical contact</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>manipulated</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

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

openly. Our output for (d) oddly misses that relief was created despite correctly identifying that 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 Márquez 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 protorole 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.

Surprisingly, we find that observed 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 annotations.

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


