An Argument-Marker Model for Syntax-Agnostic Proto-Role Labeling

Juri Opitz and Anette Frank
Department of Computational Linguistics
Heidelberg University
69120 Heidelberg, Germany
{opitz,frank}@cl.uni-heidelberg.de

Abstract

Semantic proto-role labeling (SPRL) is an alternative to semantic role labeling (SRL) that moves beyond a categorical definition of roles, following Dowty’s feature-based view of proto-roles. This theory determines agenthood vs. patienthood based on a participant’s instantiation of more or less typical agent vs. patient properties, such as, e.g., volitionality in an event. To perform SPRL, we develop an ensemble of hierarchical models with self-attention and concurrently learned predicate-argument markers. Our method is competitive with the state-of-the-art, overall outperforming previous work in two different formulations of the task (multi-label and Likert scale prediction). In contrast to previous work, our results do not depend on supplementary gold syntax.

1 Introduction

Deciding on a linguistically sound, clearly defined and broadly applicable inventory of semantic roles is a long-standing issue in linguistic theory and natural language processing. To alleviate issues found with classical thematic role inventories, Dowty (1991) argued for replacing categorical roles with a feature-based, composite notion of semantic roles, introducing the theory of semantic proto-roles (SPR). At its core, it proposes two prominent, composite role types: proto-agent and proto-patient. Proto-roles represent multi-faceted, possibly graded notions of agenthood or patienthood. For example, consider the following sentence from Bram Stoker’s Dracula:

(1) He opened it [the letter] and read it gravely.

‘Davidsonian’ analyses based on SR and SPR of the event open are displayed in Fig. 1. The SPR analysis provides more detail about the event and the roles of the involved entities. If an argument is considered an agent or patient follows from the proto-typical properties the argument exhibits: e.g., being manipulated is proto-typical for patient, while volition is proto-typical for an agent (the count is selected as agent in both events and the letter as patient).

Only recently two SPRL data sets have been published. Reisinger et al. (2015) developed a property-based proto-role annotation schema with 18 properties. A turker (selected in a pilot annotation) answered questions such as how likely is it that the argument mentioned with the verb changes location? on a 5-point Likert or responded inapplicable. This dataset (news domain) will henceforth be denoted by SPR1. Based on the experiences from the SPR1 annotation process, White et al. (2016) published SPR2 which follows a similar annotation schema, but, in contrast to SPR1, contains doubly annotated data from the web domain for 14 properties.

Our work makes the following contributions: In §2, we provide an overview of previous SPRL work and outline a common weakness: reliance on gold syntax trees. To alleviate this issue, we propose a span-based, hierarchical neural model (§3) which learns marker embeddings to highlight predicate-argument structures of events. Experiments (§4) show that our ensemble method outperforms previous works. Lastly, we perform ablation experiments to analyze the contributions of different model components.
2 Related Work

Teichert et al. (2017) (TEI17) formulate the SPRL task as a multi-label problem and develop a conditional random field model (CRF). Given an argument phrase and a corresponding predicate, the model predicts which of the 18 properties hold. Compared with a simple feature based linear model introduced by Reisinger et al. (2015) (R_EL15), the CRF exhibits superior performance by more than 10 $\Delta$macro F1. Incorporating features derived from additional gold syntax improves the CRF performance significantly. For treating the task as a multi-label problem the Likert classes $\{1, 2, 3\}$ and inapplicable are collapsed to $-$ and Likert scale classes $\{4, 5\}$ are mapped to $+$. Subsequent work, including ours, conform to this setup.

Rudinger et al. (2018) (RUD18) are the first to treat SPRL as a multivariate Likert scale regression problem. They develop a neural model whose predictions have good correlation with the values in the testing data on SPR1 and SPR2. When comparing with TEI17 (multi-label setting, SPR1), the model establishes a new state-of-the-art. Pre-training the model in a machine translation setting helps on SPR1 but results in a performance drop on SPR2. The model takes a sentence as input to a Bi-LSTM (Hochreiter and Schmidhuber, 1997) to produce a sequence of hidden states. The prediction is based on the hidden state corresponding to the head of the argument phrase, which is determined by inspection of the gold syntax tree. Our approach, in contrast, does not rely on any supplementary information from gold syntax trees.

3 Attentive Marker Model

Following previous work (RUD18), the backbone of our model is a Bi-LSTM. To ensure further comparability, pretrained 300 dimensional GloVe embeddings (Pennington et al., 2014) are used for building the input sequence $(e_1, ..., e_T)$. In contrast to RUD18, we multiply a sequence of marker embeddings $(m_1, ..., m_T)$ element-wise with the sequence of word vectors: $(e_1 \cdot m_1, ..., e_T \cdot m_T)$ (Fig. 2). We distinguish three types of marker embeddings that indicate the position of the argument in question (red, Fig. 2), the predicate (green, Fig. 2) and remaining parts of the sentence. The sequence of marked embeddings is further processed by a Bi-LSTM in order to obtain a sequence of hidden states $S = (s_1, ..., s_T)$, indicated by $\leftarrow$ and $\rightarrow$, Fig. 2).

From there, we take intuitions from Zheng et al. (2018) and compute the next sequence of vectors by letting every hidden state attend to every other hidden state, which is expressed by the following formulas:

$$h_{t,\nu} = \tanh(QS_t + KS_\nu + b_t)$$
$$e_{t,\nu} = \sigma(v^T h_{t,\nu} + \alpha_t)$$
$$a_t = \text{softmax}(e_t)$$
$$z_t = \sum_{\nu} a_{t,\nu} \cdot s_\nu$$

$Q, K$ are weight matrices, $b_t$ is a bias vector, $\alpha_t$ is a bias scalar and $v$ a weight vector. Letting every hidden state attend to every other hidden state gives the model freedom in computing the argument-predicate composition. This is desirable, since arguments and predicates frequently are in long-distance relationships (e.g., in SPR1, predicates and arguments often lie more than 10 words apart, cf. §A.3). We proceed by concatenation, $z = [z_1; ..., z_T]$, and compute intermediate outputs approximating the property-lielihood Likert scales with weight matrix $A$ and ReLU activation functions ($FF_{ReLU}$, Fig. 2):

$$a = \text{ReLU}(Az).$$

(1)

To perform the multi-label prediction with $|P|$ possible labels, we use $[a; z]$ for computing the final decisions with $2|P|$ output neurons and $|P|$ separate weight matrices ($|P|^2 FF_{ReLU}$, Fig. 2), one for each property $p \in P$:

$$o_p = \text{softmax}(W^p[a; z])$$

(2)

For the regression task, we reduce the number of output neurons from $2|P|$ to $|P|$ and use ReLU
activation functions instead. We hypothesize that the hierarchical structure can support the model in making predictions on the top layer. E.g., if the argument is predicted as most likely to not be sentient and very likely to be manipulated, the model may be less tempted to predict an awareness label at the top layer. The auxiliary loss for any data example is given as the mean square error over the auxiliary output neurons:

$$\ell' = \frac{\lambda'}{|P|} \sum_{p \in P} (a_p^* - o_p)$$

(3)

In case of the multi-label formulation, our main loss for an example is the average cross entropy loss over every property:

$$\ell = - \frac{\lambda}{|P|} \sum_{p \in P} (o^*_{p,1} \log o_{p,1} + o^*_{p,2} \log o_{p,2})$$

(4)

where $o^*_{p,0} = \mathbf{I}(\neg p)$ and $o^*_{p,1} = \mathbf{I}(p)$ i.e. the gold label indicator.

4 Experiments

We use the same data setup and split as Teichert et al. (2017); Rudinger et al. (2018) – for data and pre-processing details as well as hyper parameter choices, see these works and Appendices §A.1 & A.2. We fit an ensemble of 50 models with different random seeds on the given training data (either SPR1-train for SPR1-test predictions or SPR2-train for SPR2-test predictions), and apply early stopping on the development data (maximum Pearson’s $\rho$ for multivariate Likert regression, maximum macro F1 for the multi-label task). Unseen testing data is predicted by combining the models’ decisions in a simple majority vote when performing multi-label prediction or, when in the regression setup, by computing the mean of the output scores (for every property).

Multi-Label Prediction The results on newspaper data (SPR1) are displayed in Table 1. Our ensemble technique improves massively in the property location (+19.1 $\Delta$F1). A significant loss is experienced in the property changes possession (-9.6 $\Delta$F1). Overall, our ensemble method outperforms all prior works on SPR1 (REI15: +17.7 $\Delta$macro F1; TEI17: +6.2 $\Delta$, RUD18: +1.0 $\Delta$). To the best of our knowledge no multi-label prediction results have been published yet for SPR2.

| property | multi-label (ML), F1 score | regression (LR), $\rho$ |
|----------|--------------------------|-------------------------|
| instigated | 76.7 | 47.2 |
|ivolitional | 69.8 | 44.4 |
|awareness | 68.8 | 66.1 |
|sentient | 42.0 | 35.6 |
|exists as physical | 64.8 | 47.4 |
|created before | 79.5 | 81.7 |
|existing during | 93.1 | 71.0 |
|created after | 82.3 | 47.2 |
|destroyed | 17.1 | 18.5 |
|change | 54.0 | 35.6 |
|change state | 54.6 | 71.0 |
|change possession | 0.0 | 58.0 |
|location | 0.0 | 53.8 |
|stationary | 13.3 | 47.4 |
|physical contact | 21.5 | 47.2 |
|macr | 72.1 | 86.0 |

Table 1: SPR1 results.

As baselines we apply three label selection strategies: a majority label baseline, a constant strategy which always selects the positive label and a random baseline which samples a positive target label according to the occurrence ratio in the training data (maj, constant & ran, Table 3). The results show that our method yields massive improvements over both baselines (more than +10 $\Delta$F1) in 4 out of 14 proto role properties. For argument changes possession and awareness the improvement over both baselines is more than +25 $\Delta$F1 and for sentient more than +40 $\Delta$. However, in the partitive and for benefit properties, constant remains unbeaten by a large margin (-21.7 & -9.9 $\Delta$F1). Overall, our method yields significant improvements both over random (+27.7 $\Delta$macro F1), constant (+9.5) and majority (27.7).

Likert Scale Regression On SPR1, our model achieves large gains in correlation for the properties location\(^1\) and change of location ($\Delta \rho$: +0.136 & $\Delta \rho$: +0.066, Table 1), which is in accordance with the results for these two properties in the multi-label prediction setup. Our model is outperformed by RUD18 in the property stationary ($\Delta \rho$: -0.045). All in all, our system outperforms RUD18 ($\Delta \rho$: +0.005). On the smaller web data (SPR2), our model outperforms RUD18 slightly by +0.001$\Delta \rho$ (Table 3). However, if we compare with RUD18’s model setup which achieved the best score on the SPR1 testing data (pre-training on supervised MT task, macro regression result SPR2: 0.521$\rho$), we achieve a significantly higher macro average ($\Delta \rho$: +0.014).

\(^1\)i.e. does arg describe the location of the event?
Table 2: System properties and dependencies and results of SPRL systems. Discussion c.f. §4.

| Method   | gold syntax | superv. transfer learning | ensembling | optimistic |
|----------|-------------|----------------------------|------------|------------|
| REI15    | no          | no                         | no         | no         |
| TEI17    | yes         | yes                        | no         | yes        |
| RUD18    | yes         | yes                        | no         | yes        |
| Ours     | no          | no                         | yes        | no         |

Table 3: SPR2 results.

| property          | baselines | RUD18 | ours |
|-------------------|-----------|-------|------|
| maj ran const ours | 0.489 70.5 | 77.9 0.590 0.582 |
| volitional        | 0.391 61.8 | 88.1 0.841 0.839 |
| awareness         | 0.489 67.1 | 92.7 0.879 0.882 |
| sentient          | 0.476 44.3 | 91.9 0.880 0.874 |
| existed before    | 89.5 80.0 | 89.5 0.616 0.645 |
| existed during    | 98.0 96.2 | 97.0 0.358 0.374 |
| existed after     | 94.1 86.1 | 94.1 0.478 0.469 |
| chg state         | 0.195 31.3 | 29.7 0.352 0.351 |
| chg possession    | 0.0 5.5   | 6.6 0.488 0.520 |
| chg location      | 0.0 12.0 21.7 0.492 0.517 |
| chg state cont.   | 0.0 9.2   21.7 0.352 0.396 |
| for benefit       | 0.0 66.1 71.0 61.1 0.578 0.580 |
| was used          | 0.0 66.1 71.0 71.9 0.203 0.173 |
| partitive         | 0.0 10.4 24.2 2.5 0.359 0.283 |

Discussion
While the performance differences to RUD18 may seem marginal for many properties and overall in the regression task (Δρ SPR1: +0.005, Δρ SPR2: +0.001, Table 1 & 2), it is important to note that the result of our system has substantially fewer dependencies (Table 2): (i), our model does not rely on supplementary gold syntax – in fact, since it is span-based, our model is completely agnostic to any syntax. Besides our approach, only REI15 does not depend on supplementary gold syntax for the displayed results. However, our ensemble model outperforms REI15’s feature based linear classifier in every property (+17 Δmacro F1 in total). (ii), our results are non-optimistic: RUD18 and TEI17 evaluated multiple system configurations on the testing data and we always display the best outcome. In sum, despite having significantly less dependencies on external resources, our approach proves to be competitive with all methods from prior works.

Model Ablations
All ablation experiments are conducted in the multi-label formulation and carried out five times to assess sensitivities to different random seeds. The complete model is denoted by Full (Table 4). We proceed by ablating different components in a leave-one-out fashion: (i), SA: the self-attention components of the ensemble model are removed; (ii), markers: we abstain from highlighting the positions of arguments and predicates; (iii), hierarchy: we remove the hierarchical structure and do not predict auxiliary outputs. (iv) ensemble: only one model instead of the multi-voter ensemble.

From all ablated components, removing the markers hurts the model the most (SPR1: -17.3 Δmacro F1; SPR1: -14.5 Δ). While the self-attention component can boost the model’s performance by up to +7 ΔF1 on SPR1 and +6 ΔF1 on SPR2, the hierarchical structure leads to only marginal gains of up to +0.1 ΔF1.

5 Conclusion
In our proposed SPRL ensemble model, predicate-argument constructs are highlighted with concurrently learned marker embeddings and self-attention enables the model to capture long-distance relationships between arguments and predicates. The method overall outperforms the state-of-the-art when applied to newspaper texts (multi-label prediction macro F1: +1.0 Δ) Our model is competitive with the state-of-the-art for Likert regression on texts from the web domain and is the first to be applied for multi-label SPRL on this type of data, significantly outperforming all baselines by a large margin (+9.5 Δmacro F1; more than +40 ΔF1 for is the argument sentient?). Our SPRL model is syntax-agnostic and has fewer dependencies on supplementary gold resources, paving the way to an end-to-end SPRL system.
A Supplemental Material

A.1 Notes

Calculation of macro F1 The global performance metric for multi-label SPRL is defined as ‘macro F1’. To ensure full comparability of results, we use the same formula as prior works (Rudinger et al., 2018):

\[
\text{macro avg.} = \frac{2 \cdot P_{\text{macro avg.}} \cdot R_{\text{macro avg.}}}{P_{\text{macro avg.}} + R_{\text{macro avg.}}}
\]  

where \(P\) and \(R\) are Precision and Recall and macro avg. means the unweighted mean of these quantities computed over all proto role properties. The above macro F1 metric, though not explicitly displayed in the prior work papers, has been confirmed by the main authors (email communication).

Data split of SPR1 (Teichert et al., 2017) reframed the SPRL task as a multi-label problem. Previously the task was to answer, given a predicate and an argument, one specific proto role question (binary label or single output regression). Now we need to predict all proto role questions at once (multi-label or multi-output regression). In order to allow this formulation of the task, the authors needed to redefine the original train-dev-test split of SPR1 (recent works, including ours, all use the re-defined split).

Gold Labels We conform to prior works (Rudinger et al., 2018; Teichert et al., 2017) and (i) collapse classes in the multi-label setup from \(\{\text{NA, 1, 2, 3}\}\) and \(\{4, 5\}\) to ‘−’ and ‘+’ and (ii) treat NA as 1 in the Likert regression formulation, averaging the scores for doubly annotated data (SPR2).

Reported Numbers When comparing against prior works, we always take the best reported numbers. In the EMNLP publication of Rudinger et al. (2018), we found a few transcription errors in the result tables (confirmed by email communication with the main authors, who plan to upload an errata section). In the case of transcription errors, we took the latest reported, error-corrected numbers.

A.2 Hyperparameters & Preprocessing

The hyper parameter configuration of our model are displayed in Table 5. Further hyper-parameters include (i) handling of OOV-tokens: Look-up in pre-trained word vectors, if not in pre-trained vectors use special OOV-embedding. (ii), casing and cleaning: All tokens are modified to lower-case, strings representing floating or integer point numbers are modified such as in, e.g.: “54.2 ∆ increase” → ‘NUMBER ∆ increase’. Sequence pruning: Consider that \(I = \{i\}\) is the index of the
| hyper-parameter       | choice                                           |
|-----------------------|--------------------------------------------------|
| $\lambda$ (main loss) | 1                                                |
| $\lambda'$ (aux. losses) | 0.2                                              |
| optimizer             | Adam (Kingma and Ba, 2014)                      |
| optimizer param.      | $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e^{-6}$ |
| learning rate         | 0.001                                            |
| Bi-LSTM units         | 2 · 64                                            |
| max. seq length       | 30                                               |
| padding               | pre                                              |
| LSTM, init $U$        | random orthogonal                                |
| LSTM, init $W$        | Xavier unif. (Glorot and Bengio, 2010)           |
| FF, init              | Xavier unif.                                     |
| Trainable embeddings, init | $U(-0.05,+0.05)$                              |
| Fixed embeddings, init | GloVe 300d (Pennington et al., 2014)            |

Table 5: Hyper parameter configuration.

Figure 3: Distribution of the number of words between argument and verb (distance relationship) and sentence lengths in the data sets SPR1 and SPR2.

Predicate and $J = \{j, ..., k\}$ are the indices corresponding to the argument. As long as the input sequence length is longer than maximum length (30, cf. Table 5), we clip left tokens so that the index of the token $m$ is smaller than $\min I \cup J$, then we proceed to clip tokens to the right so that $m$ is greater than $\max I \cup J$, for the very rare cases that this was not sufficient we proceed to clip tokens with $m \notin I \cup J$ (the marker sequences are adjusted accordingly). The clipping strategy ensures that predicate and argument tokens are present in every input sequence. Sequences which are shorter than 30 words are pre-padded.

A.3 Data statistics

Statistics about the distances between arguments and predicates as well as general sentence lengths are displayed in Fig. 3.