Semantic Role Labeling with Iterative Structure Refinement

Chunchuan Lyu1 Shay B. Cohen1 Ivan Titov1,2
1ILCC, School of Informatics, University of Edinburgh
2ILLC, University of Amsterdam

chunchuan.lv@gmail.com scohen@inf.ed.ac.uk ititov@inf.ed.ac.uk

Abstract

Modern state-of-the-art Semantic Role Labeling (SRL) methods rely on expressive sentence encoders (e.g., multi-layer LSTMs) but tend to model only local (if any) interactions between individual argument labeling decisions. This contrasts with earlier work and also with the intuition that the labels of individual arguments are strongly interdependent. We model interactions between argument labeling decisions through iterative refinement. Starting with an output produced by a factorized model, we iteratively refine it using a refinement network. Instead of modeling arbitrary interactions among roles and words, we encode prior knowledge about the SRL problem by designing a restricted network architecture capturing non-local interactions. This modeling choice prevents overfitting and results in an effective model, outperforming strong factorized baseline models on all 7 CoNLL-2009 languages, and achieving state-of-the-art results on 5 of them, including English.

1 Introduction

Semantic role labeling (SRL), originally introduced by Gildea and Jurafsky (2000), involves the prediction of predicate-argument structure, i.e., identification of arguments and their assignment to underlying semantic roles. Semantic-role representations have been shown to be beneficial in many NLP applications, including question answering (Shen and Lapata, 2007), information extraction (Christensen et al., 2011) and machine translation (Marcheggiani et al., 2018). In this work, we focus on dependency-based SRL (Hajić et al., 2009), a popular version of the task which involves identifying syntactic heads of arguments rather than marking entire argument spans (see the graph in red in Figure 1). Edges in the dependency graphs are annotated with semantic roles (e.g., A0:PLEASER) and the predicates are labeled with their senses from a given sense inventory (e.g., SATISFY.01 in the example).

Before the rise of deep learning methods, the most accurate SRL methods relied on modeling high-order interactions in the output space (e.g., between arguments or arguments and predicates) (Watanabe et al., 2010; Toutanova et al., 2008). Earlier neural methods can model such output interactions through a transition system, and achieve competitive performance (Henderson et al., 2013). However, current state-of-the-art SRL systems use powerful sentence encoders (e.g., layers of LSTMs (Li et al., 2018; He et al., 2017) or multi-head self-attention (Strubell et al., 2018)) and factorize over small fragments of the predicted structures. Specifically, most modern models process individual arguments and perform predicate disambiguation independently. The trend towards more factorizable models is not unique to dependency-based SRL but common for most structured prediction tasks in NLP (Kiperwasser and Goldberg, 2016; Dozat and Manning, 2017, 2018). The only major exception is language generation tasks, especially machine translation and language modeling, where larger amounts of text are typically used in training.

Powerful encoders, in principle, can capture long-distance dependencies and hence alleviate the need for modeling high-order interactions in...
the output. However, capturing these interactions in the encoder would require substantial amounts of data. Even if we have domain knowledge about likely interactions between components of the predicted graphs, it is hard to inject this knowledge in an encoder.

Consider the example in Figure 1. The argument ‘state’ appears in the highly ambiguous syntactic position ‘[] to satisfy’. All three core semantic roles of the predicate SATISFY:01 can in principle appear here: patient (A1:ENTITY_FULFILLED, as in ‘a sweet tooth to satisfy’), instrument (A2:METHOD, as in ‘a little dessert to satisfy your sweet tooth’) and agent (A0:PLEASER, as in our actual example). The basic factorized model got it wrong, assigning A1 to the argument ‘state’. However, taking into account other arguments, the model can correct the label. The configuration ‘A1 to satisfy’ is more likely when an agent (A0) is present in the sentence. The lack of an agent boosts the score for the correct configuration ‘A0 to satisfy’.

Our iterative refinement approach encodes the above intuition. In iterative refinement (Lee et al., 2018), a refinement network repeatedly takes previous output as input and produces its refined version. Formally, we have

\[ y^{t+1} = \text{Refine}(x, y^t). \]

Naturally, such refinement strategy also requires an initial prediction \( y^0 \), which is produced by a (‘base’) factorized model.

Refinement strategies have been successful in machine translation (Lee et al., 2018; Novak et al., 2017; Xia et al., 2017; Hassan et al., 2018), but their effectiveness in other NLP tasks is yet to be demonstrated.\(^1\) We conjecture that this discrepancy is due to differences in data availability. Given larger amounts of training data typically used in machine translation, their base models and refinement networks overfit to a lesser extent. Overfitting in (1) the base model and (2) the refinement network are both problematic. The first implies that either there are no mistakes in the base models in the training set or their distribution is very different from that in the test regime, so the training material for the inference networks ends up being misleading. The second naturally means that refinement will fail at test time. We address both these issues by designing restricted inference networks and adding a specific form of noise when training them.

Our structured refinement network is simple but encodes non-local dependencies. Specifically, it takes into account the information about the role distributions on the previous iteration aggregated over the entire sentences but not the information what the other arguments are. It is a coarse compressed representation of the prediction, yet it represents long-distance information not readily available within the factorized base model. While this is not the only possible design, we believe that the empirical gains from using this simple refinement network, demonstrate the viability of our general framework of iterative refinement with restricted inference networks. They also suggest that intuitions underlying declarative constraints used in early work on SRL (Punyakanok et al., 2008; Das et al., 2012) can be revisited but now encoded in a flexible soft way to provide induction biases for the refinement networks. We leave this for future work.

We consider the CoNLL-2009 dataset (Hajič et al., 2009). We start with a strong factorized baseline model, which already achieves state-of-the-art results on a subset of the languages. Then, using our structure refinement network, we improve on this baseline on all 7 CoNLL-2009 languages. The model achieves best-reported results in 5 languages, including English. We also observe improvements on out-of-domain test sets, confirming the robustness of our approach. We perform experiments demonstrating the importance of adding noise, and ablation studies showing the necessity of incorporating output interactions. Furthermore, we provide analysis on constraint violations and errors on the English test set.\(^2\)

2 Related Work

Learning to refine predictions from neural structured prediction models has recently received significant attention. Our approach bears similarity to methods used in machine translation (Lee et al., 2018; Novak et al., 2017; Xia et al., 2017). All these methods refine a translated sentence produced by a seq2seq model with another seq2seq model. Among them, the deliberation networks

---

\(^1\)See extra discussion and related work in section 2.

\(^2\)The code and experiment settings can be accessed at https://github.com/ChunchuanLv/Iterative_Inference
by Xia et al. (2017) rely on BiLSTMs and improve initial predictions from an competitive baseline and obtain state-of-art-results on English-to-French translation. Later, it has been shown that the deliberation networks can improve translation when used within the Transformer framework (Hassan et al., 2018).

Certain approaches, not necessarily directly optimized for refinement, can nevertheless be regarded as iterative refinement methods. Structured prediction energy networks (SPENs) are trained to assign global energy scores to output structures, and the gradient descent is used during inference to minimize the global energy (Belanger and McCallum, 2016). As the gradient descent involves iterative optimization, its steps can be viewed as iterative refinement. In particular, Belanger et al. (2017) build a SPEN for SRL, but for the span-based formalism, not the dependency one we consider in this work. While they improve over their baseline model, their baseline model used multilayer perceptron to encode local factors, thus the encoder power is limited. Moreover their refined model performs worse in the out-of-domain setting than their baseline model, indicating overfitting (Belanger et al., 2017).

In the follow-up work, Tu and Gimpel (2018, 2019) introduce inference networks to replace gradient descent. Their inference networks directly refine the output. Improvements over competitive baselines are reported on part-of-speech tagging, named entity recognition and CCG supertagging (Tu and Gimpel, 2019). However, their inference networks are distilling knowledge from a tractable linear-chain conditional random field (CRF) model. Thus, these methods do not provide direct performance gains. More importantly, the interactions captured in these models are likely local, as they learn to mimic Markov CRFs.

Denoising autoencoders (Vincent et al., 2008) can also be used to refine structure. Indeed, image segmentation can be improved through iterative inference with denoising autoencoders (Romero et al., 2017; Drozdzal et al., 2018). Their framework is very similar to ours, albeit we are working in a discrete domain. One other difference is that by using a convolutional architecture in the refinement network, they are still modeling only local interactions. At a more conceptual level, Bengio et al. (2013) argued that a denoising autoencoder should not be too robust to the input variations as to ignore the input. This indicates that we should not expect refinement networks to correct all the errors, even in theory, and hence, the refinement networks do not need to be particularly powerful.

Very recently, Wang et al. (2019) used high order statistical model for Semantic Dependency Parsing (Oepen et al., 2015), and obtain improvements over strong baseline using BiLSTM. They attempted loopy belief propagation and mean field variational inference for inference, and train the model end to end. Such inference steps are well motivated. This work is similar to energy network approach (Belanger and McCallum, 2016), while a global score function is provided, and approximate inference steps are used. Comparing to ours, the inference can also be regarded as iterative structure refinement. Yet, we do not provide a global score and directly try to model the refinement. In principle, our formalization should give us more liberty in terms of designing the refinement network.

3 Dependency Semantic Role Labeling

In this section, we introduce the notation and present our factorized baseline model.

3.1 Notation

In dependency SRL, for each sentence of length \( n \), we have a sequence of words \( w \), dependency labels \( \text{dep} \), part-of-speech tags \( \text{pos} \), each being a discrete sequence of length \( n \). To simplify notation, we consider one predicate at a time. We denote the number of roles by \( r \), it includes the ‘null’ role, signifying that the corresponding word is not an argument of the predicate. Formally, \( P \in \Delta_{m-1} \) is the probability distribution over \( m \) predicate senses, and \( \Delta_{m-1} \) represents the corresponding probability simplex. We also have predicate sense embeddings \( \Pi \in \mathbb{R}^{m \times d_e} \), and index \( j \), throughout the discussion, refers to the position of the target predicate in the sentence. \( R \in \Delta_{n-1}^n \) is a matrix of size \( n \times r \) such that each row sums to 1, corresponding to a probability distribution over roles. In particular, \( R_i,0 \) is the probability of \( i \)-th word not being an argument of the predicate.

We index role label and sense predictions from different refinement iterations (‘time steps’) with \( t \), i.e. \( P^t \) and \( R^t \). The index \( t \) ranges from 0 to \( T \), and \( P^0 \) and \( R^0 \) denotes the predictions from the factorized baseline model. Details (e.g., hyperparameters) are provided in the appendix.
3.2 Factorized Model

Similarly to recent approaches to SRL and semantic graph parsing (He et al., 2017; Li et al., 2018; Dozat and Manning, 2018), our factorized baseline model starts with concatenated embeddings \( \mathbf{x} \).

Then, we encode the sentence with a BiLSTM, further extract features with an MLP (multilayer perceptron) and apply a bi-affine classifier to the resulting features to label the words with roles. We also use a predicate-dependent linear layer for sense disambiguation.

More formally, we start with getting a sentence representation by concatenating embeddings. We have \( \mathbf{x}^{w} \in \mathbb{R}^{n \times d_{w}}, \mathbf{x}^{dep} \in \mathbb{R}^{n \times d_{d}}, \mathbf{x}^{pos} \in \mathbb{R}^{n \times d_{p}} \) for words, dependency labels and part-of-speech tags, respectively. We concatenate them to form a sentence representation:

\[
\mathbf{x} = \mathbf{x}^{w} \circ \mathbf{x}^{dep} \circ \mathbf{x}^{pos} \in \mathbb{R}^{n \times d_{x}}
\]

We further encode the sentence with a BiLSTM:

\[
\mathbf{h} = \text{BiLSTM}(\mathbf{x}) \in \mathbb{R}^{n \times d_{h}}
\]

From these context-aware word representations, we produce features for argument identification and role labeling that will be used by a bi-affine classifier. Note that, for every potential predicate-argument dependency (i.e. a candidate edge), we need to produce representations of both endpoints: the argument and the predicate ‘sides’. For the argument side, \( \mathbf{h}^{\rho_{0}} \) will be used to compute the logits for argument identification and \( \mathbf{h}^{\rho_{1}} \) will be used for deciding on its role:

\[
\mathbf{h}^{\rho_{0}} = \text{MLP}(\mathbf{h}) \in \mathbb{R}^{n \times d_{\rho_{0}}}
\]

\[
\mathbf{h}^{\rho_{1}} = \text{MLP}(\mathbf{h}) \in \mathbb{R}^{n \times d_{\rho_{1}}}
\]

Similarly, for the predicate side, we also extract two representations \( \mathbf{h}^{\pi}_{0} \) and \( \mathbf{h}^{\pi}_{1} \) (recall that the predicate is at position \( j \)):

\[
\mathbf{h}^{\pi}_{0} = \text{MLP}(\mathbf{h}_{j}) \in \mathbb{R}^{d_{\rho_{0}}}
\]

\[
\mathbf{h}^{\pi}_{1} = \text{MLP}(\mathbf{h}_{j}) \in \mathbb{R}^{d_{\rho_{1}}}
\]

We then obtain logits \( I^{\rho_{0}} \) corresponding to the decision to label arguments as null, and logits \( I^{\rho_{1}} \) for other roles. So, we have:

\[
I^{\rho_{0}} = \text{BiAffine}(\mathbf{h}^{\rho_{0}}, \mathbf{h}^{\rho_{0}}) \in \mathbb{R}^{n}
\]

\[
I^{\rho_{1}} = \text{BiAffine}(\mathbf{h}^{\rho_{1}}, \mathbf{h}^{\rho_{1}}) \in \mathbb{R}^{n \times (r-1)}
\]

Unlike Dozat and Manning (2018), where argument identification and role labeling are trained with two losses,\(^3\) we feed them together into a single softmax layer to compute the semantic-role distribution \( R^{0} \):

\[
I^{\rho} = I^{\rho_{0}} \circ I^{\rho_{1}} \in \mathbb{R}^{n \times r}
\]

\[
R^{0} = \text{Softmax}(I^{\rho}) \in \Delta_{r-1}^{n}
\]

Now, for sense disambiguation, we need to extract yet another predicate representation \( \mathbf{h}^{\pi} \):

\[
\mathbf{h}^{\pi} = \text{MLP}(\mathbf{h}_{j}) \in \mathbb{R}^{d_{\pi}}
\]

In the formalism we use (PropBank), senses are predicate-specific, so we use predicate-specific sense embedding matrices \( \Pi \). The matrix \( \Pi \) acts as a linear layer before softmax:

\[
I^{\pi} = \Pi \cdot \mathbf{h}^{\pi} \in \mathbb{R}^{m}
\]

\[
I^{\rho} = \text{Softmax}(I^{\pi}) \in \Delta_{m-1}
\]

This ends the description of our baseline model, which we also use to get initial predictions for iterative refinement.

4 Structured Refinement Network

In this section, we introduce the structured refinement network for dependency SRL. When doing refinement, it has access to the roles distribution \( \mathbf{R}^{t} \in \Delta_{r-1}^{n} \) and the sense distribution \( \mathbf{P}^{t} \in \Delta_{m-1} \) computed at the previous iteration (i.e. time \( t \)). In addition, it exploits the sentence representation \( \mathbf{x} \in \mathbb{R}^{n \times d_{x}} \). Our refinement network is limited and structured, in the sense that it only has access to a compressed version of the previous prediction, and the network itself is a simple MLP.

Similarly to our baseline model, we extract feature vectors \( \mathbf{g} \) from input \( \mathbf{x} \) and further separately encode the argument representation \( \mathbf{g}^{\alpha} \) and the predicate token representation \( \mathbf{g}^{\pi} \):

\[
\mathbf{g} = \text{BiLSTM}(\mathbf{x}) \in \mathbb{R}^{n \times d_{h}}
\]

\[
\mathbf{g}^{\alpha} = \text{MLP}(\mathbf{g}) \in \mathbb{R}^{n \times d_{g}}
\]

\[
\mathbf{g}^{\pi} = \text{MLP}(\mathbf{g}) \in \mathbb{R}^{d_{g}}
\]

\(^3\)The separate processing of \( I^{\rho_{0}} \) and \( I^{\rho_{1}} \) rather than using a single MLP for all roles, including null, results in extra representation power allocated for the argument identification subtask.
4.1 Role Refinement Network

First, we describe our structured refinement network for role labeling. We use $i$ to index arguments. We obtain a compressed representation $O_i$ used for refining $R_i$ by summing up the probability mass for all roles, excluding the null role:

$$O_{i,u} = \sum_{k \neq i} R_{k,u} \in \mathbb{R}$$  \hspace{1cm} (17)

$$O_i = [O_{i,u}] \in \mathbb{R}^{r-1}$$  \hspace{1cm} (18)

Intuitively, $O_i$ is the aggregation of all other roles being labeled by the current predicate. We concatenate $O_i$ with feature vectors of the current argument $g^\alpha$, predicate $g^\pi$, the relaxed predicate sense embedding $\Pi \cdot P^\alpha$ and the role probability itself ($R_i$) to form the input to a two-layer network:

$$z_i^\alpha = R_i^\alpha \circ O_i \circ g^\alpha \circ g^\pi \circ (\Pi \cdot P^\pi)$$  \hspace{1cm} (19)

$$z_i^\alpha \in \mathbb{R}^{2r-1+2d_g+d_\pi}$$

$$M_i^\alpha = W^\alpha \cdot \sigma(W_\alpha \cdot z_i^\alpha) \in \mathbb{R}^r$$  \hspace{1cm} (20)

$$M_i^\alpha = [M_i^\alpha_{u}] \in \mathbb{R}^{n \times r}$$  \hspace{1cm} (21)

where $\sigma$ is the logistic sigmoid function, $W_\alpha \in \mathbb{R}^{d_r \times (2r-1+2d_g+d_\pi)}$, $W^\alpha \in \mathbb{R}^{r \times d_\pi}$ are learned linear mappings. We obtain our refined logits $M_i^\alpha$ for the $i$-th argument; $M^\alpha$ refers to the stacked matrix of logits for all arguments. To obtain the refined role distribution, we add up $M^\alpha$ and $1^\alpha$ that we got from the baseline model, and follow that by a softmax layer:

$$R_i^{1+} = \text{Softmax}(M^\alpha + 1^\alpha) \in \Delta_{n-1}^{r-1}$$  \hspace{1cm} (22)

4.2 Sense Refinement Network

To build a representation for sense disambiguation, we simply compute the probability mass for each role (excluding the null role) to obtain $r^\pi$, and concatenate this with $g^\pi$ and $\Pi \cdot P^\pi$:

$$r^\pi = \sum_k R_{k,1}^\pi : \in \mathbb{R}^{r-1}$$  \hspace{1cm} (23)

$$z^\pi = (\Pi \cdot P^\pi) \circ r^\pi \circ g^\pi \in \mathbb{R}^{r-1+d_g+d_\pi}$$  \hspace{1cm} (24)

Differently from the role refinement network, sense prediction is predicate-specific. Therefore, we first map $z^\pi$ to $\mathbb{R}^{d^\pi}$, and then take the inner product with the predicate-specific sense embeddings $\Pi$ to get the refined logits:

$$M^\pi = \Pi \cdot W^\pi \cdot \sigma(W^\pi \cdot z^\pi) \in \mathbb{R}^m$$  \hspace{1cm} (25)

Similarly to role refinement, $\sigma$ is the logistic function, $W^\pi \in \mathbb{R}^{d_r \times (r-1+d_g+d_\pi)}$, $W^\pi \in \mathbb{R}^{m \times d^\pi}$ are learned linear mappings. Again, we combine the logits $M^\pi$ and $1^\pi$ before the softmax layer:

$$P^{i+1} = \text{Softmax}(M^\pi + 1^\pi) \in \Delta_{m-1}$$  \hspace{1cm} (26)

4.3 Weight Tying

Our refinement networks are similar to the denoising autoencoders (DAEs; Vincent et al. 2008), so we use the weight-tying technique popular with DAEs. We believe that the technique may be even more effective here as the amount of labeled data for SRL is lower than in many usual applications of DAEs. We tie $W_\alpha$ with a subset of $W^\alpha$ rows: specifically with the rows acting on $R_i^\alpha$ in the computation of $M_i^\alpha$ (see equations 19 and 20). Similarly, we tie $W^\pi$ with the part of $W^\pi$ corresponding to $\Pi \cdot P^\pi$ (see equations 24 and 25). Formally,

$$W_\alpha = W_i^\alpha [:, r]$$  \hspace{1cm} (27)

$$W^\pi = W_i^\pi [:, d_\pi]$$  \hspace{1cm} (28)

where $W[:, k]$ takes the first $k$ rows of matrix $W$.

4.4 Self Refinement

We describe a simpler version of the refinement network which we will use in experiments to test whether the improvements with the structured refinement network over the factorized baseline are genuinely coming from modeling interaction between arguments rather than from simply combining multiple classifiers. This simpler refinement network does not account for any interactions between arguments. Instead of equations 19 and 24, we have:

$$z_i^\alpha = R_i^\alpha \circ g^\alpha \circ g^\pi \in \mathbb{R}^{2r-1+2d_g+d_\pi}$$  \hspace{1cm} (29)

$$z^\pi = (\Pi \cdot P^\pi) \circ g^\pi \in \mathbb{R}^{d_\pi}$$  \hspace{1cm} (30)

Everything else is kept the same as in the full model, expect that the size of $W_\alpha$ and $W^\pi$ needs to be adjusted. We refer to this ablated network as the self-refinement network.

5 Training for Iterative Structure Refinement

In this section, we describe our training procedure.

5.1 Two-Stage Training

We have two models: the baseline model, producing the initial predictions, and the iterative refinement network, correcting them. While it is possible to train them jointly, we find joint training slow
5.2 Stochastic Training

Our baseline model overfits to the training set, and, if simply training on its output, our refinement network would learn to copy the base predictions. Instead, we perturb the baseline prediction during training. Naturally, we can add dropout (Srivastava et al., 2014) and recurrent dropout (Gal and Ghahramani, 2016) to our neural networks. However, for the smaller data set we use, we find this not sufficient. In particular, we use Gumbel-Softmax instead of Softmax. Gumbel-Softmax(I) = Softmax(I + λgϵ), where the random variable ϵ is drawn from the standard Gumbel distribution (Maddison et al., 2017; Jang et al., 2017), and λg is a hyper-parameter controlling decoding stochasticity.\footnote{A more canonical way of controlling stochasticity is to use the temperature but we prefer not to scale the gradient.}

5.3 Loss for Iterative Refinement

Let us denote gold-standard labels for roles and predicates as R* and P*. We use two separate losses L\textsubscript{base}(R*, P*, x) and L\textsubscript{refine}(R*, P*, x) for our two-stage training. We define losses for predictions from each refinement iteration and sum them up:

\[ L\textsubscript{base}(R*, P*, x) = L(R*, R^0) + L(P*, P^0) \]

\[ L\textsubscript{refine}(R*, P*, x) = \sum_{t=1}^{T} L(R^t, R^t) + L(P^t, P^t) \]

We adopt the Softmax-Margin loss (Gimpel and Smith, 2010; Blondel et al., 2019) for individual L. Effectively, we subtract 1 from the logit of the gold label, and apply the cross entropy loss.

6 Experiments

Datasets We conduct experiments on CoNLL-2009 (Hajič et al., 2009) data set for all languages, including Catalan (Ca), Chinese (Zh), Czech (Cz), English (En), German (De), Japanese (Jp) and Spanish (Es). We use the predicted part-of-speech tags, dependency labels, and pre-identified predicate, provided with the dataset. The statistics of datasets are shown in Table 2.

Hyperparameters We use ELMo (Peters et al., 2018) for English, and FastText embeddings (Bojanowski et al., 2017; Grave et al., 2018) for all other languages. We train and run the refinement networks for two iterations. All other hyperparameters are the same for all languages, except BiLSTMs for English is larger than others.

Training Details Training the refinement network takes roughly 2 times more time than the baseline models, as it requires running BiLSTMs. The extra computation for the structured refinement network is minimal. For English, training the iterative refinement model for 1 epoch takes about 6 minutes on one 1080ti GPU. Adam is used as the optimizer (Kingma and Ba, 2015), with the learning rate of 3e-4. We use early stopping on the development set. We run 600 epochs for all baseline models, and 300 epochs for the refinement networks. Batch sizes are chosen from 32, 64, or 128 to maximize GPU memory usage. Our implementation is based on PyTorch and AllenNLP (Paszke et al., 2017; Gardner et al., 2018).

6.1 Results and Discussions

Test Results Results for all CoNLL-2009 languages on the standard (in-domain) datasets are presented in Table 1. We compare our best model to the best previous single model for the corresponding language (excluding ensemble ones). Most research has focused on English, but we include results of recent models which were evaluated on at least 3 languages. When compared to the previous models, both our models are very competitive, with the exception of German. On the German dataset, Mulcaire et al. (2018) also report a relatively weak result, when compared to Roth and Lapata (2016). The German dataset is the smallest one in terms of the number of predicates. Syntactic information used by Roth and Lapata (2016) may be very beneficial in this setting and may be the reason for this discrepancy. Our structured refinement approach improves over the best previous results on 5 out of 7 languages. Note that hyper-parameters of the refinement network are not tuned for individual languages, suggesting that the proposed method is robust and may be easy to apply to new languages and/or new base models. The only case where the refinement network was not effective is Chinese, where it achieved only a negligible improvement.

Out-of-Domain Results on the out-of-domain
<table>
<thead>
<tr>
<th>Model</th>
<th>Ca</th>
<th>Cz</th>
<th>De</th>
<th>En</th>
<th>Ja</th>
<th>Es</th>
<th>Zh</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roth and Lapata (2016)</td>
<td>-</td>
<td>-</td>
<td><strong>80.10</strong></td>
<td>-</td>
<td>-</td>
<td>80.20</td>
<td>79.4</td>
<td>-</td>
</tr>
<tr>
<td>Marcheggiani et al. (2017)</td>
<td>-</td>
<td>86.00</td>
<td>-</td>
<td>-</td>
<td>87.7</td>
<td>-</td>
<td>80.30</td>
<td>81.2</td>
</tr>
<tr>
<td>Mulcaire et al. (2018)*</td>
<td>79.45</td>
<td>85.14</td>
<td>69.97</td>
<td>87.24</td>
<td>76.00</td>
<td>77.32</td>
<td>81.89</td>
<td>79.57</td>
</tr>
<tr>
<td>Previous best single model</td>
<td>80.32</td>
<td>86.02</td>
<td><strong>80.10</strong></td>
<td>90.40</td>
<td>78.69</td>
<td>80.50</td>
<td><strong>84.30</strong></td>
<td>82.90</td>
</tr>
<tr>
<td>Baseline model</td>
<td>80.69</td>
<td>87.30</td>
<td>75.06</td>
<td>90.65</td>
<td>81.97</td>
<td>79.87</td>
<td>83.26</td>
<td>82.69</td>
</tr>
<tr>
<td>Structured refinement</td>
<td><strong>80.91</strong></td>
<td><strong>87.62</strong></td>
<td>75.87</td>
<td><strong>90.99</strong></td>
<td>82.54</td>
<td>80.53</td>
<td><strong>83.31</strong></td>
<td><strong>83.11</strong></td>
</tr>
</tbody>
</table>

Table 1: Labeled F1 score (including senses) for all languages on the CoNLL-2009 in-domain test set. For previous best result, Catalan is from Zhao et al. (2009), Japanese is from Watanabe et al. (2010), Czech is from Henderson et al. (2013), German and Spanish are from Roth and Lapata (2016), English is from Li et al. (2018) and Chinese is from Cai et al. (2018). We report the best testing results from Mulcaire et al. (2018).

<table>
<thead>
<tr>
<th></th>
<th>#sent</th>
<th>#pred</th>
<th>#pred/#sent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ca</td>
<td>13200</td>
<td>37444</td>
<td>2.84</td>
</tr>
<tr>
<td>Cz</td>
<td>38727</td>
<td>414133</td>
<td>10.69</td>
</tr>
<tr>
<td>De</td>
<td>36020</td>
<td>17400</td>
<td>0.48</td>
</tr>
<tr>
<td>En</td>
<td>39279</td>
<td>179014</td>
<td>4.56</td>
</tr>
<tr>
<td>Ja</td>
<td>4393</td>
<td>25712</td>
<td>5.85</td>
</tr>
<tr>
<td>Es</td>
<td>14329</td>
<td>43828</td>
<td>3.06</td>
</tr>
<tr>
<td>Zh</td>
<td>22277</td>
<td>102827</td>
<td>4.62</td>
</tr>
</tbody>
</table>

Table 2: Number of sentences and predicates in training set of different languages.

We observe improvements from using refinement in all the cases. This shows that our refinement approach is robust against domain shift.

**Ablations** We report development set results in different settings in Table 3. Our full model performs 2 refinement iterations, uses weight tying, and the Gumbel noise. We select the best configuration for each language to report the test set performance in Table 1 and Table 4.

As expected, weight tying is beneficial for lower-resource languages such as Catalan, Japanese and Spanish (see Table 2 for dataset characteristics). The Gumbel noise helps for all languages except for Czech and English, the two largest datasets. In particular, we observe almost no improvement on the Spanish dataset without using the Gumbel noise. We note relatively consistent but small gains from using 2 refinement iterations. The magnitude of the gains may be an artifact of us having the loss terms \( L(R^*, R_t) \) and \( L(P^*, P_t) \), encouraging not only the final (second), but also the first, iteration to produce accurate predictions. A potential alternative explanation is that our refinement network is restricted to simple interactions, resulting in the fixed point reachable in one step.

**Constraints Violation** We consider violation of unique core roles (U), continuation roles (C) and reference roles (R) constraints from Punyakanok et al. (2008); FitzGerald et al. (2015) in Table 6. U is violated if a core role (A0 - A5, AA) appears more than once; C is violated when the C-X role is not preceded by the X role (for some X); R is violated if R-X role does not appear. Our approach results in a large reduction in the uniqueness constraint violations. Our model slightly reduces the number of R violations, while He et al. (2017) reported that deterministically enforcing constraints is not helpful (albeit in span-based SRL). However learning those constraints in a soft way might be beneficial.

**Argument Interaction vs. No Argument Interaction** We compare the structured refinement network and the self-refinement network in Table 5. Both networks share the same hyperparameters. The structured refinement network consistently outperforms the self-refinement counterpart. This suggests that the refinement model benefits from accessing information about other arguments when doing refinement. In other words, modeling argument interaction appears genuinely useful.

**Improvement Decomposition** We report labeled role precision, recall and sense disambiguation accuracy in Table 7. Our structured refinement approach consistently improves over the baseline model in all metrics. While we cannot assert the improvements on all metrics are significant, this suggests that it learns some non-trivial interactions instead of merely learning to balance precision and

---

5Roth and Lapata (2016) has better in-domain testing score, but did not report the out-of-domain score.

6We set \( \lambda^*_\alpha = 5 \) for role and \( \lambda^*_\pi = 50 \) for sense, so that initial predictions contain around 20% errors.
### Table 3: Labeled F1 score (including senses) for all languages on development set for different configurations.

<table>
<thead>
<tr>
<th>Model</th>
<th>Ca</th>
<th>Cz</th>
<th>De</th>
<th>En</th>
<th>Ja</th>
<th>Es</th>
<th>Zh</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>81.69</td>
<td>88.43</td>
<td>73.97</td>
<td>89.60</td>
<td>82.96</td>
<td>80.49</td>
<td>85.27</td>
<td>83.20</td>
</tr>
<tr>
<td>Full</td>
<td><strong>82.11</strong></td>
<td>88.62</td>
<td>74.95</td>
<td>89.82</td>
<td><strong>83.60</strong></td>
<td><strong>81.19</strong></td>
<td><strong>85.52</strong></td>
<td><strong>83.69</strong></td>
</tr>
<tr>
<td>1 iteration</td>
<td>82.07</td>
<td>88.62</td>
<td><strong>75.07</strong></td>
<td>89.93</td>
<td>83.49</td>
<td>81.03</td>
<td>85.47</td>
<td>83.40</td>
</tr>
<tr>
<td>un-tied</td>
<td>81.99</td>
<td>88.61</td>
<td>75.04</td>
<td>89.79</td>
<td>83.47</td>
<td>80.89</td>
<td>85.49</td>
<td>83.61</td>
</tr>
<tr>
<td>no Gumbel</td>
<td>82.07</td>
<td><strong>88.71</strong></td>
<td>74.62</td>
<td>90.08</td>
<td>83.33</td>
<td>80.55</td>
<td>85.42</td>
<td>83.54</td>
</tr>
</tbody>
</table>

Table 4: Labeled F1 scores (including senses) on English, German, Czech in-domain and out-of-domain test sets; we chose the previous models achieving the best scores on the out-of-domain test sets.

<table>
<thead>
<tr>
<th>Language</th>
<th>Test Ood</th>
<th>Test Ood</th>
</tr>
</thead>
<tbody>
<tr>
<td>Li et al. (2018)</td>
<td>90.4</td>
<td>81.5</td>
</tr>
<tr>
<td>Baseline</td>
<td>90.65</td>
<td>81.98</td>
</tr>
<tr>
<td>Structured Refinement</td>
<td><strong>90.99</strong></td>
<td><strong>82.18</strong></td>
</tr>
<tr>
<td>Zhao et al. (2009)</td>
<td>76.19</td>
<td>67.78</td>
</tr>
<tr>
<td>Baseline</td>
<td>75.06</td>
<td>65.25</td>
</tr>
<tr>
<td>Structured Refinement</td>
<td>75.87</td>
<td>65.69</td>
</tr>
<tr>
<td>Czech</td>
<td>Test Ood</td>
<td>86.0</td>
</tr>
<tr>
<td>Marcheggiani et al. (2017)</td>
<td>87.30</td>
<td>85.80</td>
</tr>
<tr>
<td>Baseline</td>
<td>87.30</td>
<td>85.80</td>
</tr>
<tr>
<td>Structured Refinement</td>
<td><strong>87.62</strong></td>
<td><strong>86.04</strong></td>
</tr>
</tbody>
</table>

**Error Correction Analysis** We show the errors that the structured refinement network corrects in Figure 2. In the baseline confusion matrix, we see the errors are fairly balanced for all the roles we consider here. In the error correction matrix, the corrections are also fairly evenly distributed. Yet, this is not completely uniform. There is a tendency towards filtering out arguments rather than generating new ones.

### 7 Conclusions and Future Work

We propose the structured refinement network for dependency semantic role labeling. The structured refinement network corrects predictions made by a strong factorized baseline model while modeling interactions in the predicated structure. The resulting model achieves state-of-the-art results on 5 out of 7 languages in the CoNLL-2009 data set, and substantially outperforms the factorized model on all of these languages.

For the future work, the structured refinement network can be further improved. For example, we can take an inspiration from either declarative constraints used in the previous work (Punyakanok et al., 2008) or from literature on lexical semantics of verbs, studying patterns of event and argument realization (e.g., Levin 1993). Indeed, the unique role constraint as a declarative constraint is one of the motivation for the concurrent work on modeling argument interaction in SRL (Chen et al., 2019). That work relies on capsule networks (Sabour et al., 2017) and focuses primarily on enforcing the role uniqueness constraint.

The framework can be extended to other tasks. For example, in syntactic dependency parsing: the refinement network can rely on representations of grandparent nodes, siblings and children to propose a correction. In general, structure refinement networks should allow domain experts to incorporate prior knowledge about output dependencies and improve model performance.

### Acknowledgments

We thank the anonymous reviewers for their suggestions. The project was supported by the European Research Council (ERC StG BroadSem...
Table 5: Labeled F1 score (including senses) for all languages on the CoNLL-2009 in-domain test set.

<table>
<thead>
<tr>
<th>Model</th>
<th>Ca</th>
<th>Cz</th>
<th>De</th>
<th>En</th>
<th>Ja</th>
<th>Es</th>
<th>Zh</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline Model</td>
<td>80.69</td>
<td>87.30</td>
<td>75.06</td>
<td>90.65</td>
<td>81.97</td>
<td>79.87</td>
<td>83.26</td>
<td>82.69</td>
</tr>
<tr>
<td>Self Refinement</td>
<td>80.65</td>
<td>87.32</td>
<td>74.83</td>
<td>90.71</td>
<td>82.27</td>
<td>80.08</td>
<td><strong>83.32</strong></td>
<td>82.74</td>
</tr>
<tr>
<td>Structured Refinement</td>
<td><strong>80.91</strong></td>
<td><strong>87.47</strong></td>
<td><strong>75.83</strong></td>
<td><strong>90.83</strong></td>
<td><strong>82.54</strong></td>
<td><strong>80.53</strong></td>
<td><strong>83.31</strong></td>
<td><strong>83.06</strong></td>
</tr>
</tbody>
</table>

Table 6: Unique core roles violations (U), continuation roles violations (C) and reference roles violations (R) on English in-domain test set.

<table>
<thead>
<tr>
<th>Model</th>
<th>U</th>
<th>C</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gold</td>
<td>55</td>
<td>0</td>
<td>88</td>
</tr>
<tr>
<td>Baseline</td>
<td>301</td>
<td>2</td>
<td>114</td>
</tr>
<tr>
<td>Structured Refinement</td>
<td>142</td>
<td>2</td>
<td>111</td>
</tr>
</tbody>
</table>

Table 7: Labeled roles precision (RP), recall (RR) and sense disambiguation accuracy (Sense) on English in-domain test set.

<table>
<thead>
<tr>
<th>Model</th>
<th>RP</th>
<th>RR</th>
<th>Sense</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>88.1</td>
<td>88.3</td>
<td>96.2</td>
</tr>
<tr>
<td>Structured Refinement</td>
<td>88.7</td>
<td>88.5</td>
<td>96.3</td>
</tr>
</tbody>
</table>

References


