

# Embedded-State Latent Conditional Random Fields for Sequence Labeling

Dung Thai Sree Harsha Ramesh Shikhar Murty Luke Vilnis Andrew McCallum

College of Information and Computer Sciences

University of Massachusetts Amherst

{dthai, shramesh, smurty, luke, mccallum}@cs.umass.edu

## Abstract

Complex textual information extraction tasks are often posed as sequence labeling or *shallow parsing*, where fields are extracted using local labels made consistent through probabilistic inference in a graphical model with constrained transitions. Recently, it has become common to locally parametrize these models using rich features extracted by recurrent neural networks (such as LSTM), while enforcing consistent outputs through a simple linear-chain model, representing Markovian dependencies between successive labels. However, the simple graphical model structure belies the often complex non-local constraints between output labels. For example, many fields, such as a first name, can only occur a fixed number of times, or in the presence of other fields. While RNNs have provided increasingly powerful context-aware local features for sequence tagging, they have yet to be integrated with a global graphical model of similar expressivity in the output distribution. Our model goes beyond the linear chain CRF to incorporate multiple hidden states per output label, but parametrizes their transitions parsimoniously with low-rank log-potential scoring matrices, effectively learning an embedding space for hidden states. This augmented latent space of inference variables complements the rich feature representation of the RNN, and allows exact global inference obeying complex, learned non-local output constraints. We experiment with several datasets and show that the model outperforms baseline CRF+RNN models when global output constraints are necessary at inference-time, and explore the interpretable latent structure.

## 1 Introduction

As with many other prediction tasks involving complex structured outputs, such as image segmentation (Chen et al., 2018), machine transla-

tion (Bahdanau et al., 2015), and speech recognition (Hinton et al., 2012), deep neural networks (DNNs) for sequence labeling and shallow parsing have become standard tools for information extraction (Collobert et al., 2011; Lample et al., 2016). In the language of structured prediction, DNNs process the input sequence to produce a rich *local* parametrization for the output prediction model. However, output variables obey a variety of hard and soft constraints — for example, in sequence tagging tasks such as named entity recognition, I-PER cannot follow B-ORG.

Interestingly, even with such powerful local featurization, the DNN model does not automatically capture a mode of the output distribution through local decisions alone, and can violate these constraints. Successful applications of DNNs to sequence tagging gain from incorporating a simple linear chain probabilistic graphical model to enforce consistent output predictions (Collobert et al., 2011; Lample et al., 2016), and more generally the addition of a graphical model to enforce output label consistency is common practice for other tasks such as image segmentation (Chen et al., 2018).

Previous work in DNN-featurized sequence tagging with graphical models for information extraction has limited its output structure modeling to these simple local Markovian dependencies. In this work, we explore the addition of latent variables to the prediction model, and through a parsimonious factorized parameter structure, perform representation learning of hidden state embeddings in the graphical model, complementary to the standard practice of representation learning in the local potentials of the segmentation model. By factorizing the log-potentials of the hidden state transition matrices, we are able to learn large numbers of hidden states without overfitting, while the latent dynamics add the capability to learn global

```

**GROUND TRUTH**
0 P/BED . Candice P/FIRSTNAME Kosgei P/LASTNAME , 0 yrs old P/AGE
under Dr Harrod P/DOC_LASTNAME , came in with cyctic fibrosis P/DIAGNOSIS .
She P/GENDER s still the same MYSHIFT_STATUS . No other changes . All obs are
all normal MYSHIFT/OTHER .

**CRF**
0 P/BED . Candice P/FIRSTNAME Kosgei P/LASTNAME , 0 yrs old P/AGE
under Dr Harrod P/DOC_LASTNAME , came in with cyctic fibrosis P/DIAGNOSIS .
She P/GENDER s still the same MYSHIFT_STATUS . No other changes . All obs are
all normal MYSHIFT_STATUS .

**OUR MODEL**
0 P/BED . Candice P/FIRSTNAME Kosgei P/LASTNAME , 0 yrs old P/AGE
under Dr Harrod P/DOC_LASTNAME , came in with cyctic fibrosis P/DIAGNOSIS .
She P/GENDER s still the same MYSHIFT_STATUS . No other changes . All obs are
all normal MYSHIFT/OTHER .

```

Figure 1: An example result from the CLEF eHealth dataset. The soft output constraint suggests tagging patient status as *Myshift/Others* if there already is a *Myshift\_Status* tag. Note that we have the same phrase tagged as *Myshift\_Status* in the training dataset.

constraints on the overall prediction, without sacrificing efficient exact inference.

While soft and hard global constraints have a rich history in sequence tagging (Koo et al., 2010; Rush and Collins, 2012; Anzaroot et al., 2014), they have been underexplored in the context of neural-network based feature extraction models. In response, we present a latent-variable CRF model with a novel mechanism for learning latent constraints without overfitting, using low-rank embeddings of large-cardinality latent variables. For example, these non-local constraints appear in fine-grained nested field extraction, which requires hierarchical consistency between the subtags of an entity. Further, information extraction and slot filling tasks often require domain specific constraints — for example, we must avoid extracting the same field multiple times. A good combination of input featurization and output modeling is needed to capture these structural dependencies.

In this work we present a method for sequence labeling in which representation learning is applied not only to inputs, but also to output space, in the form of a lightly parameterized transition function between a large number of latent states. We introduce a hidden state variable and learn the model dynamics in the hidden state space rather than the label state space. This relaxes the Markov

assumption between output labels and allows the model to learn global constraints. To avoid the quadratic blowup in parameters with the size of the latent state space, we factorize the transition log-potentials into a low-rank matrix, avoiding overfitting by effectively learning parsimonious embedded representations of the latent states. While the low rank *log*-potential matrix does not improve test-time inference speed, we can perform exact Viterbi inference to compute the labeling sequence. Figure 1 shows an example where our model finds the correct labeling sequence while a standard DNN+CRF model fails, by obeying a global constraint learned from the training data.

We examine the performance of the Embedded-State Latent CRF on two datasets: citation extraction on the UMass Citations dataset and medical record field extraction on the CLEF dataset. We observe improved performance in both tasks, whose outputs obey complex structural dependencies that are not able to be captured by RNN featurization. Our biggest improvement comes in the medical domain, where the small training set gives our parsimonious approach to output representation learning an extra advantage.

## 2 Proposed Model

### 2.1 Problem Formulation

We consider the sequence labeling task, defined as follows. Given an input text sequence with  $T$  tokens  $\mathbf{x} = \{x_1, x_2, \dots, x_T\}$ , find a corresponding output sequence  $\mathbf{y} = \{y_1, y_2, \dots, y_T\}$  where each output symbol  $y_i$  is one of  $N$  possible output labels. There are structural dependencies between the output labels, and resolving such dependencies is necessary for good performance.

### 2.2 Background

The input featurization in our model is similar to previously mentioned existing methods for tagging with DNNs (Collobert et al., 2011). We represent each input token  $x_t$  with a word embedding  $w_t$ . We then feed the embedded sequence  $\mathbf{w} = \{w_1, w_2, \dots, w_T\}$  into a bidirectional LSTM (Graves and Schmidhuber, 2005). As a result, each input  $x_t$  is associated with a contextualized feature vector  $f_t = [\vec{f}_t; \overleftarrow{f}_t] \in \mathbb{R}^d$  where  $\vec{f}_t$  and  $\overleftarrow{f}_t$  represent the left and right context at time step  $t$  of the sequence.

In this work, we concern ourselves with the mapping from these input features to a distribution

over output label sequences.

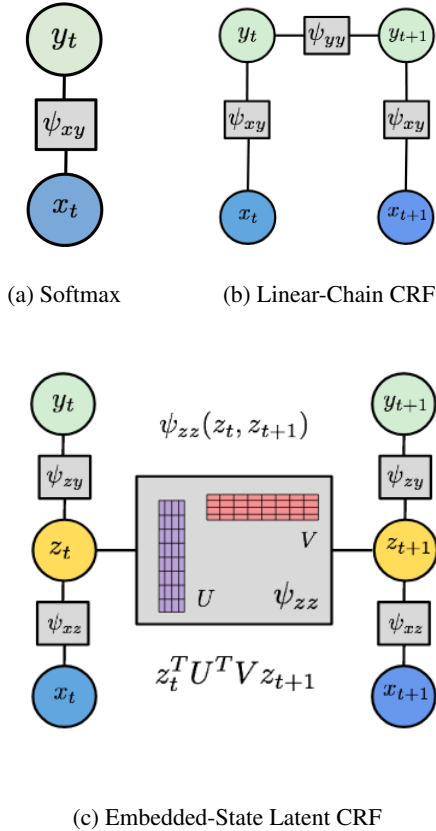


Figure 2: Comparing PGMs for tag prediction.

A straightforward solution is to use a feed-forward network to map the feature vector to the corresponding label. From a probabilistic perspective, this method is equivalent to the probabilistic graphical model in Fig.2a. Here, the goal is to estimate the posterior distribution:

$$\mathbb{P}(\mathbf{y} | \mathbf{x}) = \prod_{i=1}^T P(y_t | x_t) = \prod_{i=1}^T \psi(y_t; x_t) \quad (1)$$

where the joint distribution over the sequence is fully factorized, i.e. there is no structural dependency between  $y_t$  and the distribution  $P(y_t | x_t)$  is parameterized by a deep neural network  $\psi(y_t; x_t) = \text{feed\_forward}(f_t, y_t)$ . This model ignores all the structural dependencies between the output labels during prediction, though not featurization, and has been found unsuitable for structured prediction tasks on sequences (Collobert et al., 2011).

In order to enforce some local output consistency, Collobert et al. (2011) introduce a linear chain Conditional Random Field (CRF) layer to the model (Fig.2b). They define the energy func-

tion for a particular configuration as follows

$$\mathcal{E}(\mathbf{y} | \mathbf{x}) = \sum_{t=1}^T \psi_{xy}(x_t, y_t) + \psi_{yy}(y_t, y_{t+1}) \quad (2)$$

where the local log-potentials  $\psi_{xy}$  are parameterized by a DNN, and (for their application) the edge log-potentials  $\psi_{yy}$  are parameterized by an input-independent parameter matrix, modeling the intra-state dependencies under a Markovian assumption, giving the data log-likelihood as

$$\log \mathbb{P}(\mathbf{y} | \mathbf{x}) = \mathcal{E}(\mathbf{y} | \mathbf{x}) - \log \sum_{\mathbf{y}'} \exp(\mathcal{E}(\mathbf{y}' | \mathbf{x})) \quad (3)$$

Collobert et al. (2011) show a +1.71 performance gain in Named-Entity Recognition (NER) by explicitly enforcing these local structural dependencies. However, the Markov assumption is limiting, and much of the gain comes from enforcing deterministic hard constraints of the segmentation encoding (e.g. *I-ORG* cannot go after *B-PER*). Similar types of local gains come from hierarchical tagging schemes (e.g. *I-DATE* should be tagged as *I-VENUE/DATE* if it appears inside the *I-VENUE/\** segment). We would like to model, and learn, global, semantically meaningful soft constraints, e.g. *BOOKTITLE* should become *TITLE* if another *TITLE* does not appear in the same citation (Anzaroot et al., 2014). The state transition dynamics of the linear-chain CRF model are limited by a restriction to interaction between  $N$  output labels. The information-rich features  $f_t \in \mathbb{R}^d$  input to the local potential are restricted to a local preference over the  $N$  labels in output space, failing to exploit the full power of the underlying feature space.

### 2.3 Embedded-State Latent CRF

Our proposed model, the embedded-state latent CRF, is shown in Figure 2c. We introduce a sequence of hidden states  $\mathbf{z} = \{z_1, z_2, \dots, z_T\}$  where  $z_t$  is one of  $M$  possible discrete hidden states and  $M \gg N$ . Similarly, the corresponding energy for a particular joint configuration over  $\mathbf{y}$  and  $\mathbf{z}$  is

$$\mathcal{E}(\mathbf{y}, \mathbf{z} | \mathbf{x}) = \sum_{t=1}^T \psi_{xz}(x_t, z_t) + \psi_{yz}(y_t, z_t) + \psi_{zz}(z_t, z_{t+1}) \quad (4)$$

where  $\psi_{xz}(x_t, z_t)$ ,  $\psi_{yz}(y_t, z_t)$  are the local interaction log-potentials between the input features

and hidden states, and the hidden states and output states, respectively. The hidden state dynamics come from the log-scores  $\psi_{zz}(z_t, z_{t+1})$  for transitioning between hidden state  $z_t$  to  $z_{t+1}$ . The posterior distribution over output labels can be computed by summing over all possible configurations of  $\mathbf{z}$

$$\mathbb{P}(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z} \sum_{\mathbf{z}} \exp(\mathcal{E}(\mathbf{y}, \mathbf{z} \mid \mathbf{x})) \quad (5)$$

where  $Z = \sum_{\mathbf{y}'} \sum_{\mathbf{z}'} \exp(\mathcal{E}(\mathbf{y}', \mathbf{z}' \mid \mathbf{x}))$  is the partition function. The local log-potentials  $\psi_{xz}(x_t, z_t)$  are produced by an affine transform from the RNN feature extractor, and the output potentials  $\psi_{zy}(z_t, y_t)$  are many-to-one mappings from the hidden state, with learned potentials but pre-allocated numbers of states for each output label.

**Factorized transition log-potentials** We empirically observe that introducing a large number of hidden states can lead to overfitting, due to over-parameterization of the output dependencies. For example, *JOURNAL* often co-occurs with *PAGES* but *JOURNAL* is not strictly accompanied by *PAGES* (Anzaroot et al., 2014). Therefore, we regularize the state transition log-potential with a low-rank constraint, forming an embedding matrix wherein state transition interaction scores are mediated through low-dimensional *state embeddings* rather than a fully unconstrained parameter matrix. Instead of learning  $A \in \mathbb{R}^{M \times M}$ , a full-rank hidden state transition potential, we learn a low-rank model  $A = U^T V$  where  $U$  and  $V$  are two *rank-k* matrices. This reduces the number of parameters from  $M^2$  to  $2Mk$  (where  $k \ll M$ ) and shares statistical strength when learning transitions between similar states.

**Inference.** The brute-force computation of the posterior distribution using (5) is intractable, especially with the large number of hidden states. Fortunately, both the energy and the partition function can be computed efficiently using tree belief propagation. Due to the deterministic mapping from hidden states to outputs, we can simply fold the local input *and* output potentials  $\psi_{xz}(x_t, z_t)$  and  $\psi_{yz}(y_t, z_t)$  into the edge potentials and perform the forward-backward algorithm as in a standard linear-chain CRF. This deterministic mapping also lets us enforce hard transition constraints while retaining exact inference. Furthermore, since our implementation is in PyTorch

(Paszke et al., 2017), we only need to implement the forward pass, as automatic differentiation (back-propagation) is equivalent to the backward pass (Eisner, 2016).

**MAP inference.** At test time, we run the Viterbi algorithm to search for the best configuration over  $\mathbf{z}$  rather than over  $\mathbf{y}$ . Mapping from the hidden state  $z_t$  to the output label  $y_t$  is deterministic given the output state embedding.

### 3 Related Work

Much deep learning research concerns itself with learning to represent the structure of input space in a way that is highly predictive of the output. In this work, while using state-of-the-art sequence tagging baselines for input representation learning, we concern ourselves with learning the global structure of the output space of label sequences, as well as fine-grained local distinctions in output space. While representation learning in the form of fine-grained, discrete, latent state transitions in the output space has been explored in this context (e.g. various latent-variable conditional random fields (Quattoni et al., 2007; Sutton et al., 2007; Morency et al., 2007) and latent structured support vector machines (Yu and Joachims, 2009)), we enable the use of many more hidden states without overfitting by factorizing the log-potential transition matrices and modeling the log-scores of latent state interactions as products of low-dimensional embeddings, effectively performing feature learning in output space.

A simple linear-chain CRF over the labels was used in early applications of deep learning to sequence tagging (Collobert et al., 2011), as well as the most recent high-performing segmentation models for named entity recognition (Lample et al., 2016). Outside of NLP, in tasks such as computer vision, certain classes of fully-connected graphical models over the output pixels have been used for multi-dimensional smoothing (Adams et al., 2010; Krähenbühl and Koltun, 2011), borrowing techniques for the graphics literature.

However, none of these models performs representation learning in the output space, as in the case of our proposed embedded latent-state model. Srikumar and Manning (2014) propose a similar factorized representation of output labels and their transitions, but only apply this to pairwise transitions of output labels and not latent dynamics of the whole sequence, while we believe the biggest

gains are to be found by marrying representation learning techniques with latent variable methods.

In the graphical models literature, the most similar work to ours is the Latent-Dynamic CRF of [Morency et al. \(2007\)](#), who propose the same graphical model structure, without the deep input featurization, or more importantly, the learned embedded factorization of transition scores. Additionally, that work uses a deterministic mapping of equal numbers of hidden states to output labels, while we have a hard-constrained (hidden states to output variables are always many-to-one), but learned, potential with different outputs pre-allocated different numbers of states based on corpus frequency.

Many graphical models have been proposed for natural language processing under hard and soft global constraints, e.g. ([Koo et al., 2010](#); [Anzaroot et al., 2014](#); [Vilnis et al., 2015](#)), many based on dual decomposition ([Rush and Collins, 2012](#)). However, the constraints are often fixed, and even when learned ([Anzaroot et al., 2014](#); [Vilnis et al., 2015](#)), the learning is done simply on constraint weights generated from pre-made templates, the construction of which requires domain knowledge.

Finally, Structured Prediction Energy Networks ([Belanger and McCallum, 2016](#); [Belanger et al., 2017](#)) have been used for NLP tasks such as semantic role labeling, but they perform approximate inference through gradient descent on a learned energy function over labelings, effectively a fully-connected graphical model, while our model sits more clearly within the framework of graphical models, permitting exact inference with only nonconvex learning, common to all latent-variable models.

## 4 Experiments

We experiment on two datasets with a rich output label space, the UMass Citations dataset ([Anzaroot and McCallum, 2013](#)) and the CLEF eHealth dataset ([Suominen et al., 2015](#)). Both of the datasets have a hierarchical label space, enforced by hard transition constraints, making this a form of shallow parsing ([Anzaroot et al., 2014](#)), with additional soft constraints in the label space due to the interdependent nature of the fields being extracted.

## 4.1 Datasets

### 4.1.1 UMass Citations

We experiment with citation field extraction on the UMass Citations dataset ([Anzaroot and McCallum, 2013](#)), a collection of 2476 richly labeled citation strings, each tagged in a hierarchical manner, across a set of 38 entities demarcating both coarse-grained labeled segments, such as title, date, authors and venue, as well as fine-grained inner segments where applicable. The data follows a train/dev/test split of 1454, 655 and 367 citations, with 231085 total tokens. For example, a person’s last name could be tagged as `AUTHORS/PERSON/PERSON-LAST` or `VENUE/EDITOR/PERSON/PERSON-LAST` depending on whether the person is the author of the cited `TITLE` or an editor of the publication `VENUE`. Similarly, year could be tagged as either `DATE/YEAR` or `VENUE/DATE/YEAR` depending on whether it is the cited work’s publication date or the publication date of the venue of the cited work.

### 4.1.2 CLEF eHealth

We perform our second set of sequence labeling experiments on the NICTA Synthetic Nursing Handover dataset ([Suominen et al., 2015](#)) for clinical information extraction, consisting of 101 documents totaling 32122 tokens.

It is a synthetic dataset of handover records, which contain patient profiles as written by a registered nurse (RN) working in the medical ward and delivering verbal handovers to another nurse at a shift change by the patients bedside. A document is typically 100-300 words long, and the included handover information contains five coarse entities i.e., `PATIENTINTRODUCTION`, `MYSHIFT`, `APPOINTMENTS`, `MEDICATION` and `FUTURECARE`. Similar to the setup of the citation field extraction task described in Section 4.1.1, each of these coarse categories has a further level of nested finer labels and the entities to be identified are all hierarchical in nature. For example, the `PATIENTINTRODUCTION` section contains entities such as `PATIENTINTRODUCTION/LASTNAME` and `PATIENTINTRODUCTION/UNDERDR_LASTNAME`, the `APPOINTMENTS` section contains `APPOINTMENT/PROCEDURE_CLINICIANLASTNAME`, and `MEDICATION` contains `MEDICATION/DOSAGE` and `MEDICATION/MEDICINE`. There are a total of 35 such fine-grained entities. In addition to

the hard-constrained hierarchical structure of the labels, the task also exhibits interesting global constraints, such as only tagging the first occurrence of the patient’s gender, or the convention of labeling the most brief description of a nurse’s shift status as MYSHIFT/STATUS, while the details of the shift are tagged as MYSHIFT/OTHER. In such cases, information extraction benefits from modeling output label dependencies, as we show in the results section.

## 4.2 Training Details

Our baseline is the BiLSTM+CRF model from Lample et al. (2016), employing a bidirectional LSTM with 500 hidden units for input featurization to capture long-range dependencies in the input space. Since we do not focus on input featurization, we do not use character-level embeddings in the baseline model.

Both the baseline model and our EL-CRF model were implemented in PyTorch. For training our models, we use the hyper-parameter settings from the LSTM+CRF model of Lample et al. (2016). Although, we did explore different optimizer techniques to enhance SGD such as Adam (Kingma and Ba, 2015), Averaged SGD (Polyak and Juditsky, 1992) and YellowFin (Zhang et al., 2017), none of them performed as well as mini-batch SGD with a batch-size of 1. We also employed gradient clipping to a norm of 5.0, a learning rate of 0.01, learning rate decay of 0.05, dropout with  $p = 0.5$ , and early stopping, tuned on the citation development data. We initialized our word level embeddings using pre-trained 100 dimensional Glove embeddings (Pennington et al., 2014), which gave better performance on our tasks than the skip-n-gram embeddings (Ling et al., 2015) used in the original work of Lample et al. (2016). The datasets were pre-processed to zero-replace all occurrences of numbers. Finally, we experimented with both IOBES and IOB tagging schemes, with IOB demonstrating higher performance on our tasks.

**Embedding size** We tune the embedding size (rank constraint) for the hidden state matrix  $A$ , varying from 10 to 40, alongside the neural network parameters, and report results when fixing the other hyperparameters and varying embedding size, similar to ablation analysis. Table 4 shows the impact of different embedding sizes on the performance of the model. We found that a size of

20 works best for both datasets, confirming the importance of the rank-constrained log-potential when using large-cardinality hidden variables.

**Mapping tags to hidden states** We find that the mapping between tags and hidden states greatly influences the performance of the model. We experimented with several heuristics (e.g., individual IOB tag count ratio and entity count ratio), and found that allocating a number of hidden states proportional to the entity count gives us the best performance.

### 4.2.1 Evaluation

We report field-level F1 scores as computed using the `conlleval.pl` script.

Since the train/validation/test splits were clearly defined for the UMass Citation dataset, we trained the models on the training split, tuned the hyperparameters on the validation split and report the scores on the test dataset. However, as there were only 101 documents in the CLEF eHealth dataset, we report the Leave-One-Out (LOO) cross-validation F1 scores for this dataset i.e., we trained 101 models each with a different held-out document, merged the respective test outputs, and computed the F1 score on this merged output.

## 4.3 Results

Table 1 shows that overall performance on the UMass Citation dataset using the embedded-state latent CRF (95.18) is marginally better than the baseline BiLSTM+CRF model (95.07). However, examining the entities with the largest F1 score improvement in Table 2, we see that they are mostly within the VENUE section, which has long-range constraints with other sections, giving evidence of the model’s ability to learn constraints from the citation dataset.

Dataset	CRF	EL-CRF	+
UMASS CITATION	95.07	95.18	<b>0.11</b>
CLEF EHEALTH	68.66	70.32	<b>1.66</b>

Table 1: Entity-level F1 scores of the embedded-state latent CRF and BiLSTM+CRF baseline.

Table 1 demonstrates that EL-CRF outperforms the BiLSTM+CRF on both datasets, with larger gains on the much smaller CLEF data. Table 3 shows the top-gaining entities include MEDICATION\_MEDICINE and MEDICATION\_DOSAGE,

Label	CRF	EL-CRF	+	S
V/DEPARTMENT	66.67	100	<b>33.33</b>	1
V/STATUS	77.78	87.5	<b>9.72</b>	9
V/E/P/ PERSON_MIDDLE	83.33	91.67	<b>8.34</b>	11
REFERENCE_ID	85.11	93.02	<b>7.91</b>	22
V/SERIES	55.17	61.54	<b>6.37</b>	12
V/ADDRESS	78.85	84.31	<b>5.46</b>	46

Table 2: Top 5 entities in terms of F1 improvement on the UMass Citation Dataset. The column *S* shows the support for a given entity in the test dataset. Key for contracted entity names: V: VENUE, E: EDITOR, P: PERSON

Label	CRF	EL-CRF	+	S
P/DR/GIVENNAMES/ INITIALS	33.33	64.29	<b>30.96</b>	15
A/PROCEDURE/ TIME	34.78	53.66	<b>18.88</b>	28
M/MEDICINE	55.28	71.1	<b>15.82</b>	157
FA/WARNING/ ABNORMALRESULT	0	11.43	<b>11.43</b>	59
M/DOSAGE	9.09	18.75	<b>9.66</b>	37

Table 3: Top 5 entities in terms of F1 improvement on the CLEF eHealth dataset. Key for contracted entity names: P/DR: PATIENT\_INTRODUCTION/UNDER\_DR, A: APPOINTMENT, M: MEDICATION, FA: FUTURE\_ALERT

due to the global constraint that those entities always co-occur.

## 5 Qualitative Analysis

In this section, we provide qualitative evidence that the embedded-state latent CRF learns constraints which are not captured by the standard CRF.

First, we pick a few representative examples from the UMass Citations dataset and discuss when our model is able to correctly determine the label sequence based on the output constraints. In addition to the the hard constraints arising from hierarchical segmentation, this dataset also exhibits empirical pairwise constraints between fields e.g. two different authors’ first names cannot be placed next to each other. Figure 3 demonstrates that the CRF model fails to enforce such constraints.

Another constraint we observe in the citation data

Factor Size	UMass Citation	CLEF eHealth
10	94.92	70.06
<b>20</b>	<b>95.18</b>	<b>71.51</b>
30	94.91	69.92
40	94.88	70.33
Full Rank	95.13	71.11

Table 4: Comparison of F1 scores obtained by varying the *factor\_size* parameter, and setting the other model and neural network parameters from the model with the best cross validation.

**\*\*GROUND TRUTH\*\***

[Ma] REF-MARKER Yu . AUTHORS/FIRSTNAME I. AUTHORS/MIDDLE Manin ,  
AUTHORS/LASTNAME Cubic Forms , TITLE

**\*\*CRF\*\***

[Ma] REF-MARKER Yu . AUTHORS/FIRSTNAME I. AUTHORS/FIRSTNAME Manin ,  
AUTHORS/LASTNAME Cubic Forms , TITLE

**\*\*OUR\*\***

[Ma] REF-MARKER Yu . AUTHORS/FIRSTNAME I. AUTHORS/MIDDLE Manin ,  
AUTHORS/LASTNAME Cubic Forms , TITLE

Figure 3: Authors name constraint violation

is that the *Venue/Series* tag only appears once per citation if *Venue/Booktitle* is also present. Our model obeys this constraint and marks the whole span as *Title* instead of breaking it into *Title* and *Venue/Series*, even though the input text for that segment in isolation could represent a valid series (Figure4).

**\*\*GROUND TRUTH\*\***

Macdonald , AUTHORS/LASTNAME G. AUTHORS/FIRSTNAME J. F. AUTHORS/MIDDLE  
(0000) . DATE/YEAR The Rotation of the Earth , A Geophys- ical Discussion . TITLE  
Cambridge Univ VENUE/PUBLISHER

**\*\*CRF\*\***

Macdonald , AUTHORS/LASTNAME G. AUTHORS/FIRSTNAME J. F. AUTHORS/MIDDLE  
(0000) . DATE/YEAR The Rotation of the Earth , TITLE A Geophys- ical Discussion  
. VENUE/SERIES Cambridge Univ VENUE/PUBLISHER

**\*\*OUR\*\***

Macdonald , AUTHORS/LASTNAME G. AUTHORS/FIRSTNAME J. F. AUTHORS/MIDDLE  
(0000) . DATE/YEAR The Rotation of the Earth , A Geophys- ical Discussion . TITLE  
Cambridge Univ VENUE/PUBLISHER

Figure 4: Title should not co-occur with series.

Sometimes output structural dependencies are not able to resolve ambiguity in the labeling sequence. In Figure 5 our model correctly predicts the presence of a *Venue/Booktitle* and a *Venue/Series*, but it fails to correctly assign the entity labels.

```

**GROUND TRUTH**
0000 , VENUE/DATE/YEAR Astronomical Society of the Pacific Conference Series ,
VENUE/BOOKTITLE Vol . 000 , VENUE/VOLUME Astronomical Data Analysis Software
and Systems XI , VENUE/SERIES D. VENUE/EDITOR/FIRSTNAME
**CRF**
0000 , VENUE/DATE/YEAR Astronomical Society of the Pacific Conference Series ,
VENUE/SERIES Vol . 000 , VENUE/VOLUME Astronomical Data Analysis Software and
Systems XI , VENUE/SERIES D. VENUE/EDITOR/FIRSTNAME
**OUR**
0000 , VENUE/DATE/YEAR Astronomical Society of the Pacific Conference Series ,
VENUE/SERIES Vol . 000 , VENUE/VOLUME Astronomical Data Analysis Software and
Systems XI , VENUE/BOOKTITLE D. VENUE/EDITOR/FIRSTNAME

```

Figure 5: There is at most one series per citation.

The CLEF eHealth dataset holds a different set of constraints than the citation data, and its input sequences are not strong local indicators of the labeling sequence. Therefore, our model shows stronger performance over the Markovian baseline for this dataset. Some of the constraints concern the number of entities per document. For example, we only tag the first occurrence of a gender indicator e.g. *he*, *she*, *her*, etc., or the most general status of a nurse’s shift.

```

**GROUND TRUTH**
Alex PI/FIRSTNAME Nasser PI/LASTNAME , 0 PI/AGE under DrMorris
PI/DOC_LASTNAME . He PI/GENDER came in with back pain and fever and query
urinary stones PI/DIAGNOSIS . He is afebrile MYSHIFT/OTHER on my shift , on same
MEDICATION_DOSAGE regimen MEDICATION_MEDICINE , no other changes and the
doctors has n't come in to see him today . He is all stable MYSHIFT_STATUS .
**CRF**
Alex PI/FIRSTNAME Nasser PI/LASTNAME , 0 PI/AGE under DrMorris
PI/DOC_LASTNAME . He PI/GENDER came in with back pain and fever PI/DIAGNOSIS
and query urinary stones PI/DIAGNOSIS . He PI/GENDER is afebrile MYSHIFT/OTHER
on my shift , on same regimen MYSHIFT_STATUS , no other changes and the doctors
has n't come in to see him today . He is all stable MYSHIFT_STATUS .
**OUR MODEL**
Alex PI/FIRSTNAME Nasser PI/LASTNAME , 0 PI/AGE under DrMorris
PI/DOC_LASTNAME . He PI/GENDER came in with back pain and fever and query
urinary stones PI/DIAGNOSIS . He is afebrile MYSHIFT/OTHER on my shift , on same
MEDICATION_DOSAGE regimen MEDICATION_MEDICINE , no other changes and the
doctors has n't come in to see him today . He is all stable MYSHIFT_STATUS .

```

Figure 6: The gender indicator constraint and nurse’s shift status constraint in the CLEF eHealth dataset.

Finally, a T-SNE (Maaten and Hinton, 2008) clustering on the embedding vectors of the output tags, shown in Figure 7, demonstrates that output

structural dependencies can be reflected in tag embedding space.

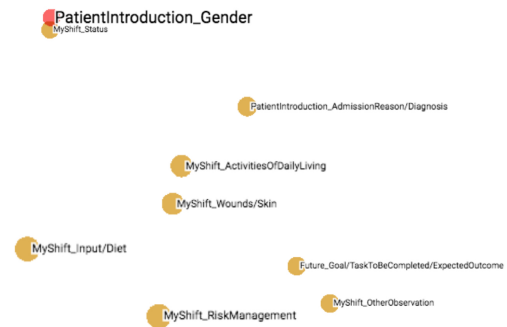


Figure 7: A part of the T-SNE clustering of the tag embedding from the CLEF eHealth dataset. The two tag PATIENTINTRODUCTION\_GENDER and MYSHIFT\_STATUS are under the similar constraint of being tagged only once per document.

## 6 Conclusion & Future Work

We present a latent variable model which not only parametrizes local potentials with the learned features from a deep neural network, but learns embedded representations in a large hidden state space, leveraging feature learning in both the input and output representations. Experimental results demonstrate that our model can learn global structural dependencies in the presence of ambiguities that cannot be resolved by local featurization of the input sequence. We find interpretable structure in the output state embeddings.

Future work will apply our model to larger datasets with more complex dependencies, and introduce multiple latent states per time-step, enabling exponentially more expressivity in output states at the cost of exact inference. We will also explore approximate inference methods, such as expectation propagation, to speed up message passing in the regime of low-rank log-potentials.

## References

- Andrew Adams, Jongmin Baek, and Myers Abraham Davis. 2010. Fast high-dimensional filtering using the permutohedral lattice. In *Computer Graphics Forum*, volume 29, pages 753–762. Wiley Online Library.
- Sam Anzaroot and Andrew McCallum. 2013. A new dataset for fine-grained citation field extraction.



- Sam Anzaroot, Alexandre Passos, David Belanger, and Andrew McCallum. 2014. Learning soft linear constraints with application to citation field extraction. *arXiv preprint arXiv:1403.1349*.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *ICLR*.
- David Belanger and Andrew McCallum. 2016. Structured prediction energy networks. In *International Conference on Machine Learning*, pages 983–992.
- David Belanger, Bishan Yang, and Andrew McCallum. 2017. End-to-end learning for structured prediction energy networks. *ICML*.
- Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 2018. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *PAMI*.
- Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.
- Jason Eisner. 2016. Inside-outside and forward-backward algorithms are just backprop (tutorial paper). In *Proceedings of the Workshop on Structured Prediction for NLP*, pages 1–17.
- Alex Graves and Jürgen Schmidhuber. 2005. Frame-wise phoneme classification with bidirectional lstm networks. In *Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on*, volume 4, pages 2047–2052. IEEE.
- Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97.
- Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *ICLR*.
- Terry Koo, Alexander M Rush, Michael Collins, Tommi Jaakkola, and David Sontag. 2010. Dual decomposition for parsing with non-projective head automata. In *EMNLP*.
- Philipp Krähenbühl and Vladlen Koltun. 2011. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*, pages 109–117.
- Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *ACL*.
- Wang Ling, Yulia Tsvetkov, Silvio Amir, Ramon Fernandez, Chris Dyer, Alan W Black, Isabel Trancoso, and Chu-Cheng Lin. 2015. Not all contexts are created equal: Better word representations with variable attention. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1367–1372.
- Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.
- Louis-Philippe Morency, Ariadna Quattoni, and Trevor Darrell. 2007. Latent-dynamic discriminative models for continuous gesture recognition. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE.
- Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. In *NIPS-W*.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.
- Boris T Polyak and Anatoli B Juditsky. 1992. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855.
- Ariadna Quattoni, Sybor Wang, Louis-Philippe Morency, Morency Collins, and Trevor Darrell. 2007. Hidden conditional random fields. *IEEE transactions on pattern analysis and machine intelligence*, 29(10).
- Alexander M Rush and MJ Collins. 2012. A tutorial on dual decomposition and lagrangian relaxation for inference in natural language processing. *Journal of Artificial Intelligence Research*, 45:305–362.
- Vivek Srikumar and Christopher D Manning. 2014. Learning distributed representations for structured output prediction. In *Advances in Neural Information Processing Systems*, pages 3266–3274.
- Hanna Suominen, Liyuan Zhou, Leif Hanlen, and Gabriela Ferraro. 2015. Benchmarking clinical speech recognition and information extraction: new data, methods, and evaluations. *JMIR medical informatics*, 3(2).
- Charles Sutton, Andrew McCallum, and Khashayar Rohanimanesh. 2007. Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. *Journal of Machine Learning Research*, 8(Mar):693–723.
- Luke Vilnis, David Belanger, Daniel Sheldon, and Andrew McCallum. 2015. Bethe projections for non-local inference. *arXiv preprint arXiv:1503.01397*.

Chun-Nam John Yu and Thorsten Joachims. 2009. Learning structural svms with latent variables. In *Proceedings of the 26th annual international conference on machine learning*, pages 1169–1176. ACM.

Jian Zhang, Ioannis Mitliagkas, and Christopher Ré. 2017. Yellowfin and the art of momentum tuning. *arXiv preprint arXiv:1706.03471*.