Learning Distributed Representations for Structured Output Prediction

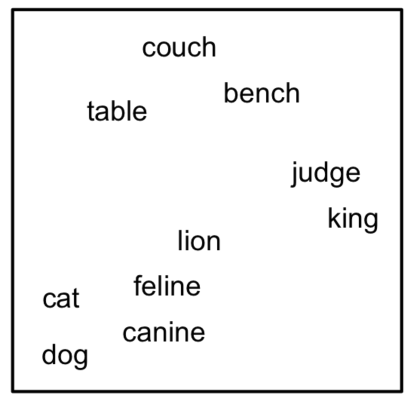
Vivek Srikumar and Christopher D. Manning

Key Points

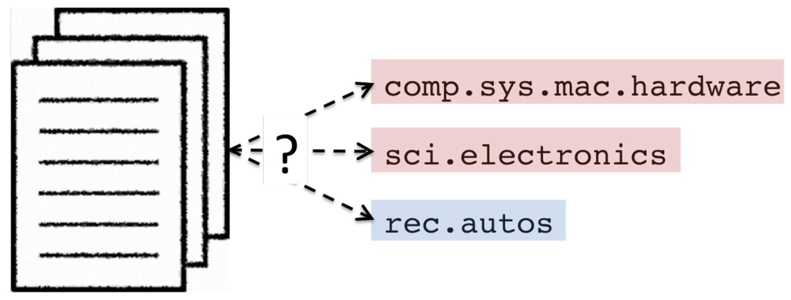
Distributed representations for inputs: A successful idea

Superficially different inputs may share meaning. E.g. Words are not discrete units of meaning.

Distributed representations allow sharing of statistical information across inputs.
E.g. vector representations for words.



Observation: Predicted labels are not discrete units of meaning

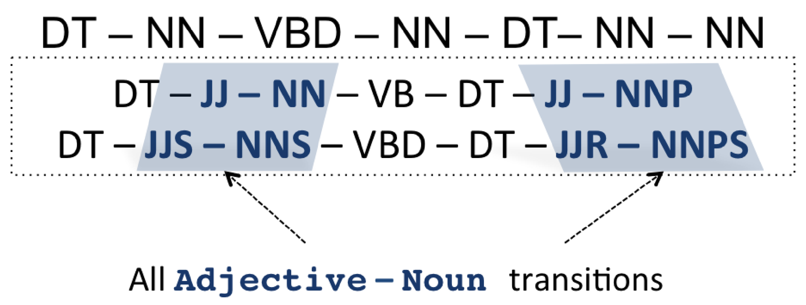


Labels encode rich semantic information with varying degrees of similarities to each other. Yet, standard models treat them as separate, discrete objects!

Structures are not discrete units of meaning

Graphs labeled with semantically rich labels are not discrete objects either. Some graphs are closer in meaning than others.

All three sequences here are equidistant by Hamming distance.



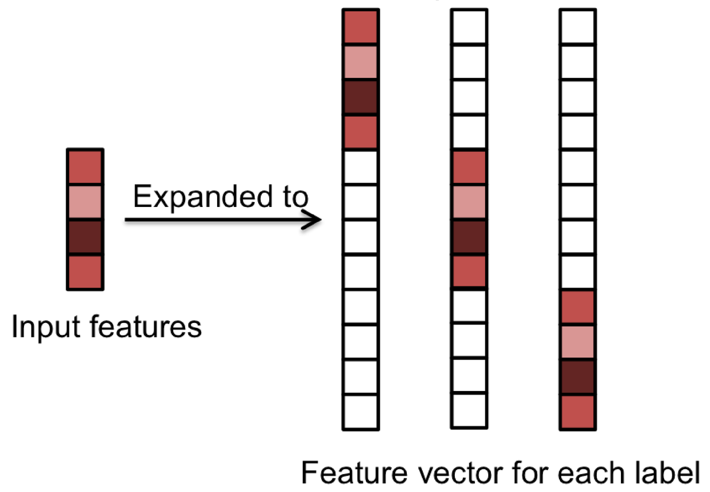
This paper: Distributed representations for structured output

The Setup: Standard Structured Models

(i.e conditional random fields, structural support vector machines)

Goal: Score structures (represented as feature vectors) to find the highest scoring one.

- Multiclass classification (i.e **atomic output**)



By construction, separate part of the parameter vector associated with each label.

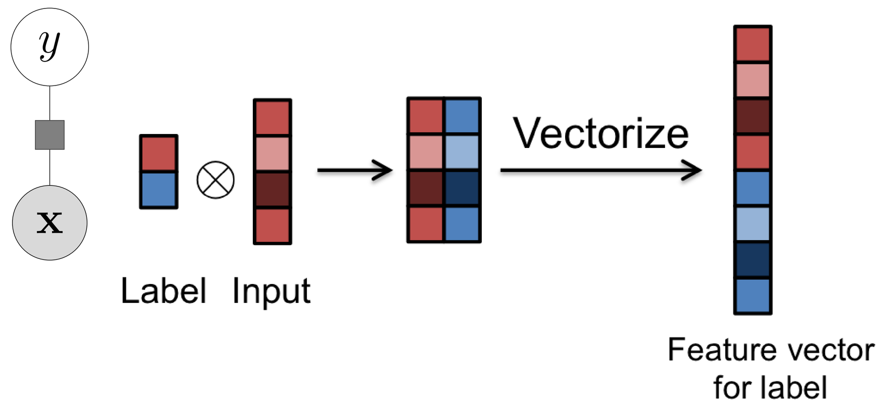
Labels do not interact with each other!

- Same applies to **compositions** of atomic labels
  E.g. transitions of sequence models, where pairs of labels are assigned different weights
- Complex structures (like sequences) can have both atomic and compositional parts.
  The feature vector for a structure simply sums over features of all the parts.
- No sharing of information across vaguely defined labels and their compositions.

DIStributed for Structured Output (DISTRO)

1. Represent atomic labels by dense vectors

Define $\Phi_{\mathbf{A}}(\mathbf{x}, y) = vec\left(\mathbf{a}_y \otimes \phi(\mathbf{x})\right)$

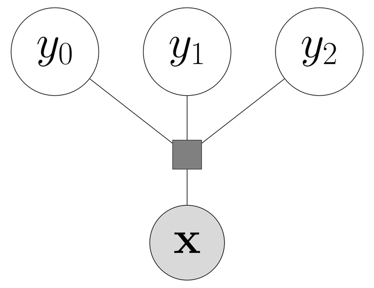


Standard multi-class classification = one-hot encoding of labels

Lower dimensional and dense label vectors force the labels' feature vectors to interact.

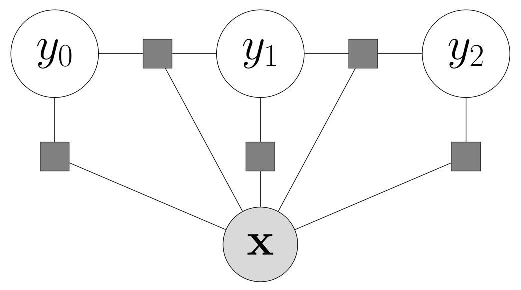
2. Composing labels via tensor products

$$\Phi_{\mathbf{A}}(\mathbf{x}, \mathbf{y}) = vec\left(\mathbf{a}_{l_{y_0}} \otimes \mathbf{a}_{l_{y_1}} \otimes \mathbf{a}_{l_{y_2}} \otimes \phi_p(\mathbf{x})\right)$$



Generalize from multiclass by defining feature vectors via the tensor product of vectors corresponding to its components.
See paper for the recursive definition.
One-hot encoding gives us standard conditional models.

3. General structures



General structures are a collection of atomic and compositional parts. The feature vector for a structure sums over the features for its parts.

Comments

1. Feature vectors redefined. Given label vectors, scoring of structures and inference is same as usual.
2. Dimensionality of label vectors is a parameter to the problem. Lower dimensionality encourages parameters to be shared across labels.

Learning

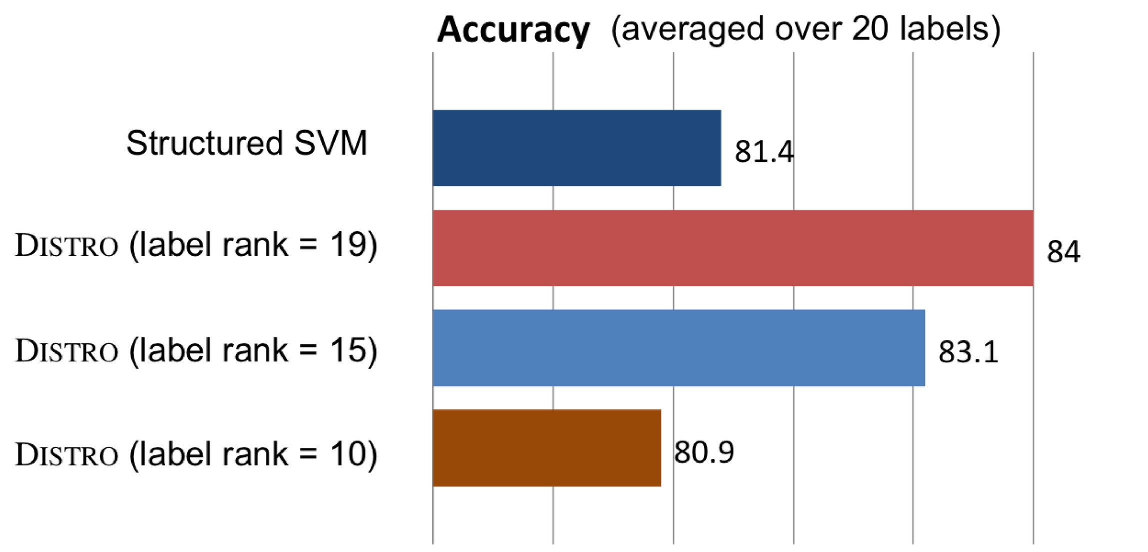
1: Initialize $\mathbf{A}^0$ randomly
2: Initialize $\mathbf{w}^0 = \min_{\mathbf{w}} f(\mathbf{w}, \mathbf{A}^0)$
3: **for** $t = 1, \cdots, T$ **do**
4: $\quad \mathbf{A}^t \leftarrow \min_{\mathbf{A}} f(\mathbf{w}^{t-1}, \mathbf{A})$
5: $\quad \mathbf{w}^{t+1} \leftarrow \min_{\mathbf{w}} f(\mathbf{w}, \mathbf{A}^t)$
6: **end for**
7: **return** $(\mathbf{w}^{T+1}, \mathbf{A}^T)$

Objective is not convex in both $\mathbf{A}$ and $\mathbf{w}$. Alternating algorithm to learn the parameters.
In experiments, $L =$ structured hinge loss.

Experiments

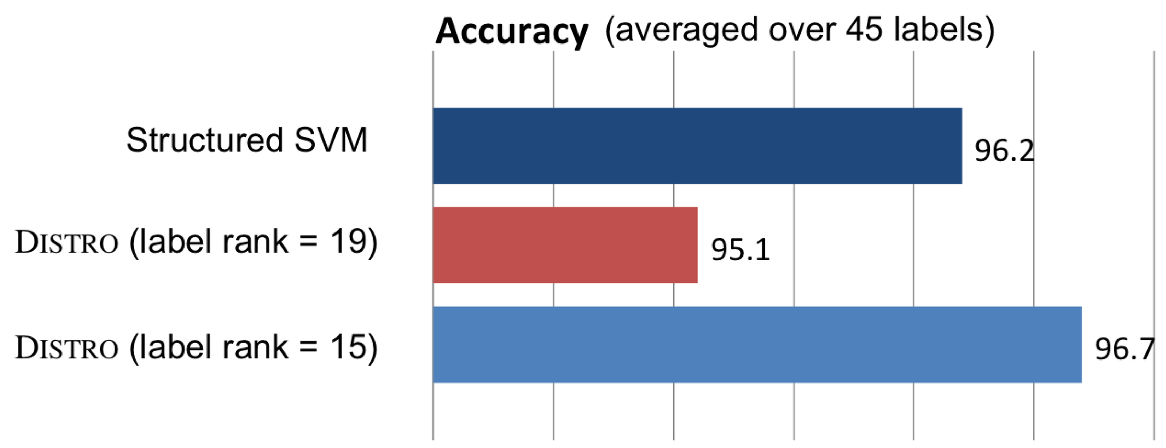
Document classification

20 newsgroup classification: Multiclass problem with semantically rich labels



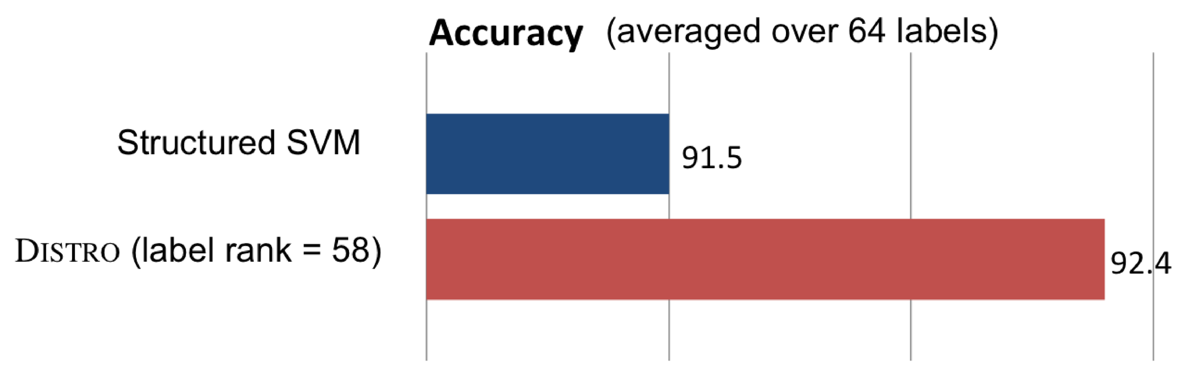
English part-of-speech classification

First order sequence model, English Penn Treebank tag set



Basque part-of-speech classification

First order sequence model, labels themselves are defined to be compositional



Final words

Related

- Embedding inputs: [Turian, et al. 2010], [Collobert, et al. 2011], [Mikolov, et al. 2013], [Coates et al. 2011], $\cdots$
- Multiclass + matrix factorization: [Srebro, et al. 2004], [Abernethy, et al. 2006]
- Connectionist++ models: [Hinton 1988], [Smolensky 1990], [Plate 1995]

This work: Arbitrary structures

- Represent atomic labels by dense vectors
- Use tensor products to construct compositional structures
- Generalizes standard CRF/structured SVM