Probing a Semantic Dependency Parser for Translational Relation Embeddings

Riley Capshaw Marco Kuhlmann Eva Blomqvist

Department of Computer and Information Science
Linköping University, Sweden

LINKÖPING UNIVERSITY

Outline

Abstract

In order to assess whether Translational Relation Embedding models are compatible with the NLP task of Semantic Dependency Parsing, we present a series of probing experiments.

We show that there seems to be some compatibility, but that further work is needed to take advantage of it (i.e., such a model is not explicitly learned by the parser).

We hope that this can be used to improve the compatibility between components in pipelines for Knowledge Graph generation and completion.

Motivation

- Translational models provide explicit edge/label embeddings.
- Improve the interpretability of the parser's label decisions.
- Improve compatibility of a parser with machine readers and other NLP/NLU pipelines (harmonize representations)

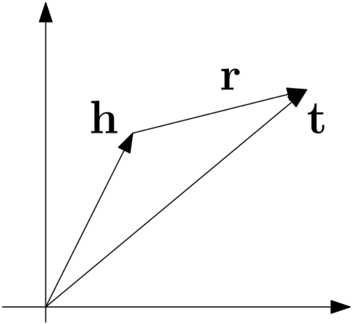
Translational Relation Embeddings (TransE)

Given facts as triples of the form (s, p, o), represent the entities s, o as vectors $\mathbf{h}, \mathbf{t} \in \mathbb{R}^n$.

Map p to the translation vector $\mathbf{r}$ such that

$$\mathbf{r} \approx \mathbf{t} - \mathbf{h}$$

TransE:



TransE: Bordes, Antoine, et al. "Translating embeddings for modeling multi-relational data." Advances in neural information processing systems. 2013.
Image: Wang, Zhen, et al. "Knowledge graph embedding by translating on hyperplanes." Twenty-Eighth AAAI conference on artificial intelligence. 2014.

Semantic Dependency Parsing

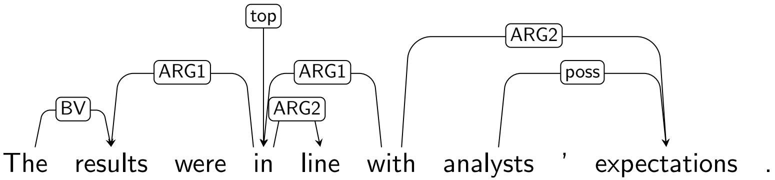
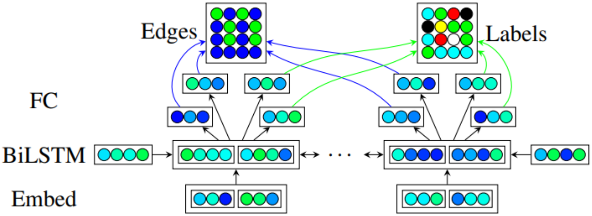


Figure: Example SDG #22007003 from the DM dataset.

- Encode shallow semantic phenomena between words.
- Formulated as a directed acyclic graph.
- Example triple: (in, `ARG1`, results)

Semantic Dependency Parsing

Semantic Dependency Parsers generally use deep neural networks in an encoder-decoder model:



- Encoder: 3-Layer of BiLSTM units
- Decoder: Biaffine classifiers

Figure: Dozat, Timothy, and Christopher D. Manning. "Simpler but more accurate semantic dependency parsing." 2018.

Probing Neural Networks

Use unintrusive simple classifiers to “study” what a neural network learns:

- Linear probes: How easily can features at a given layer linearly separate the target classes?
- Structural probes: How well can the features at a given layer be structured in a particular way?

Probing Neural Networks: Structural probes

Given a pre-trained encoder (contextualizer) for a semantic dependency parser, we want to see how well a translational model would work as the parser's decoder *without* further training the parser.

That is, we want to see if we can *explain* the parser's predictions based on a linear restructuring of its encoder's vector space.

Probing Neural Networks: Structural probes

For a given layer of a neural network:

- Take that layer's output for words i, j as the vectors $\mathbf{x}_i, \mathbf{x}_j$.
- Train a translational relation model to predict the predicted output relation for the samples given only $\mathbf{x}_i$ and $\mathbf{x}_j$ as input.
- Interpret the accuracy as a measure of the ability of the model to explain the parser's predictions.

Note that structural probes tend to probe for a feature external to the dataset, such as syntactic trees in language models. We are looking for relational structuring in a parser's predictions, which is arguably external.

Probing Neural Networks: Linear probes

To understand how well the structural probes fit the vector space, we use linear probes as a control to provide a *theoretical upper bound*.

For a given layer of a neural network:

- Combine that layer's output for words *i* and *j* as the vector $\mathbf{x}$.
- Train a simple linear softmax classifier $\text{softmax}(\mathbf{W}\mathbf{x} + \mathbf{b})$ by minimizing cross-entropy to predict the predicted[1] output label for words *i* and *j* given only their representations.
- Interpret the accuracy as a measure of the linear separability of the layer's features.

[1] Most probes predict the *true* output label, *not* the predicted one.

Parser and Data

- We used a pre-existing semantic dependency parser[2].
- We trained the parser on the English DELPH-IN MRS (DM) portion of the 2014 and 2015 SemEval shared tasks on broad-coverage semantic dependency parsing.
- For every *predicted* dependency r between pairs of words h and t, we record the respective m-dimensional representations $\mathbf{h}, \mathbf{t} \in \mathbb{R}^m$ generated by each BiLSTM layer as $(\mathbf{h}, r, \mathbf{t})$.
- Predictions are made for every sentence in both the training and testing sets.

[2] Kurtz, R., Roxbo, D., Kuhlmann, M. "Improving semantic dependency parsing with syntactic features." Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing. 2019.

Probes

Given the predicted training and testing sets, we trained the following three linear probes:

- $\mathbf{W}[\mathbf{t} - \mathbf{h}] + \mathbf{b}$ (subtraction/translation)
- $\mathbf{W}[\mathbf{h} + \mathbf{t}] + \mathbf{b}$ (addition)
- $\mathbf{W}[\mathbf{h}; \mathbf{t}] + \mathbf{b}$ (concatenation)

And the following two structural probes:

- $\mathbf{h} + \mathbf{r}_r - \mathbf{t}$
- $\mathbf{Mh} + \mathbf{r}_r - \mathbf{Mt}$

Structural Probes

For the structural probes, we first define a scoring function:

$$f_r(\mathbf{h}, \mathbf{t}) = \|\mathbf{h} + \mathbf{r}_r - \mathbf{t}\|_2^2 \text{ , or}$$

$$f_r(\mathbf{h}, \mathbf{t}) = \|\mathbf{M}\mathbf{h} + \mathbf{r}_r - \mathbf{M}\mathbf{t}\|_2^2$$

Where $\mathbf{M}$ (if present) and $\mathbf{r}_r$ are learned parameters.
Then, given a margin γ, we define our learning constraints:

1. $f_r(\mathbf{h}, \mathbf{t}) \leq \gamma$ (relaxed translation)
2. $f_{r'}(\mathbf{h}, \mathbf{t}) > \gamma$ for all $r' \neq r$ (label separation)
3. $\|\mathbf{t} - \mathbf{h}\|_2^2 > 2\gamma$ (enforce directionality)

where $\|\cdot\|_2^2$ is the squared ℓ_2 norm.

Structural Probes

Combining the constraints and scoring function, we get the following loss function:

$$\mathcal{L} = \left[f_r\left(\mathbf{h},\mathbf{t}\right) - \gamma\right]_+ + \sum_{r' \neq r}\left[\gamma - f_{r'}\left(\mathbf{h},\mathbf{t}\right)\right]_+ + \left[2\gamma - \|\mathbf{r}\|_2^2\right]_+ ,$$

where $[\cdot]_+$ is equivalent to $\max\left(0,\cdot\right)$.

We also recalculated scores where a prediction is considered correct only if the margin constraint $\left(f_r\left(\mathbf{h},\mathbf{t}\right) \leq \gamma\right)$ was satisfied.

Results: Layer-by-layer

Category	ID	Probe	Layer 0	Layer 1	Layer 2	Layer 3
Linear	L_1	$\mathbf{W}[\mathbf{h}+\mathbf{t}]+\mathbf{b}$	66.73	81.34	88.85	91.01
	L_2	$\mathbf{W}[\mathbf{t}-\mathbf{h}]+\mathbf{b}$	67.82	85.18	94.07	95.89
	L_3	$\mathbf{W}[\mathbf{h};\mathbf{t}]+\mathbf{b}$	72.69	89.06	96.54	97.52
Structural	S_1	$\mathbf{h}+\mathbf{r}_r-\mathbf{t}$	38.87	48.44	56.73	60.09
	S_2	$\mathbf{Mh}+\mathbf{r}_r-\mathbf{Mt}$	60.37	76.88	86.96	90.76
Constrained	C_1	$\mathbf{h}+\mathbf{r}_r-\mathbf{t}$	0.82	0.89	0.84	0.04
	C_2	$\mathbf{Mh}+\mathbf{r}_r-\mathbf{Mt}$	35.62	56.37	65.10	73.57

Scores for all experiments in terms of recall as a function of layer. In general, as the depth increases, the score increases.

Results: Final Layer

Category	ID	Probe	Layer 3
Linear	L_1	$\mathbf{W}[\mathbf{h}+\mathbf{t}]+\mathbf{b}$	91.01
	L_2	$\mathbf{W}[\mathbf{t}-\mathbf{h}]+\mathbf{b}$	95.89
	L_3	$\mathbf{W}[\mathbf{h};\mathbf{t}]+\mathbf{b}$	97.52
Structural	S_1	$\mathbf{h}+\mathbf{r}_r-\mathbf{t}$	60.09
	S_2	$\mathbf{Mh}+\mathbf{r}_r-\mathbf{Mt}$	90.76
Constrained	C_1	$\mathbf{h}+\mathbf{r}_r-\mathbf{t}$	0.04
	C_2	$\mathbf{Mh}+\mathbf{r}_r-\mathbf{Mt}$	73.57

Scores for all probing experiments in terms of recall.

Results: Linear Probes

Category	ID	Probe	Layer 3
Linear	L_1	$\mathbf{W}[\mathbf{h}+\mathbf{t}]+\mathbf{b}$	91.01
	L_2	$\mathbf{W}[\mathbf{t}-\mathbf{h}]+\mathbf{b}$	95.89
	L_3	$\mathbf{W}[\mathbf{h};\mathbf{t}]+\mathbf{b}$	97.52

Observations:

- Concatenation (preservation of all features) performed best.
- Translation somewhat worse, but contains most needed information.
- Addition much worse.

Results: Structural Probes

Category	ID	Probe	Layer 3
Structural	S_1	$\mathbf{h} + \mathbf{r}_r - \mathbf{t}$	60.09
	S_2	$\mathbf{Mh} + \mathbf{r}_r - \mathbf{Mt}$	90.76
Constrained	C_1	$\mathbf{h} + \mathbf{r}_r - \mathbf{t}$	0.04
	C_2	$\mathbf{Mh} + \mathbf{r}_r - \mathbf{Mt}$	73.57

Constrained means that a correct prediction is considered incorrect at prediction time if the constraints are not all satisfied.

Results: Structural Probes

Category	ID	Probe	Layer 3
Structural	S_1	$\mathbf{h} + \mathbf{r}_r - \mathbf{t}$	60.09
	S_2	$\mathbf{Mh} + \mathbf{r}_r - \mathbf{Mt}$	90.76
Constrained	C_1	$\mathbf{h} + \mathbf{r}_r - \mathbf{t}$	0.04
	C_2	$\mathbf{Mh} + \mathbf{r}_r - \mathbf{Mt}$	73.57

Observations:

- S_2 performs roughly on par with the linear probes.
- C_1 implies that S_1 largely fell back to linear classification, which implies *no* explicit structuring learned by the parser.
- C_2 performed decently well. May imply that there is some latent structure learned by the parser.

Conclusions and Future Work

- The parser does *not* explicity learn a translational relation model.
- The parser may implicitly learn such a model, or one may be easily learned from its contextualized word representations.
- This implies a compatibility with such a model, which would yield *useful* relation embeddings.
- Such embeddings could, in future work, be used to enhance end-to-end neural pipelines, such as knowledge graph generation systems or machine readers.

Thank you!