Do Transformer Networks Improve the Discovery of Rules from Text?

Mahdi Rahimi   Mihai Surdeanu

Computational Language Understanding Lab at the University of Arizona

Abstract

With their Discovery of Inference Rules from Text (DIRT) algorithm, Lin and Pantel (2001) made a seminal contribution to the field of rule acquisition from text, by adapting the distributional hypothesis of Harris (1954) to patterns that model binary relations such as *X treat Y*, where patterns are implemented as syntactic dependency paths. DIRT's relevance is renewed in today's neural era given the recent focus on interpretability in the field of natural language processing. We propose a novel take on the DIRT algorithm, where we implement the distributional hypothesis using the contextualized embeddings provided by BERT, a transformer-network-based language model (Vaswani et al., 2017; Devlin et al., 2018). In particular, we change the similarity measure between pairs of slots (i.e., the set of words matched by a pattern) from the original formula that relies on lexical items to a formula computed using contextualized embeddings. We empirically demonstrate that this new similarity method yields a better implementation of the distributional hypothesis, and this, in turn, yields patterns that outperform the original algorithm in the question answering-based evaluation proposed by Lin and Pantel (2001).

Motivation

- Lin and Pantel (2001) proposed a method called DIRT for the acquisition of rules (a.k.a patterns or paths) from text.
- In this work, we aim to combine the advantages of modern neural directions with the benefits provided by the rule-based DIRT method.

Review of DIRT

- DIRT learns inference rules from text such as *"X is the author of Y ≈ X writes Y"*
- Some of these inferences are not exact paraphrases (but are still relevant and potentially useful) such as *"X is the author of Y ≈ X is known for Y"*
- In order to do that, DIRT extends Harris' Distributional Hypothesis principle to rules:

> "If two patterns tend to link the same sets of words, we hypothesize that their meanings are similar."

Three Stages of DIRT:

1. Extraction of syntactic paths

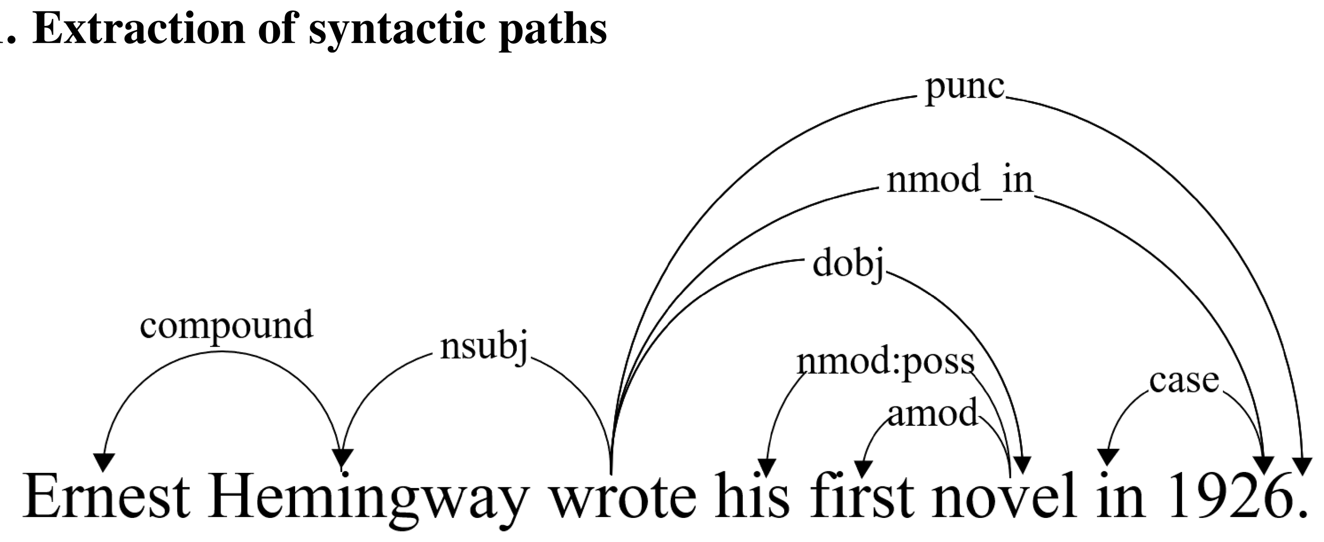

Figure 1: Example dependency tree, which matches the syntactic path `nsubj`←write→`dobj`.

2. Computing similarity between two paths

$$mi(p, slot, w) = log(\frac{|p, slot, w| \times |*, slot, *|}{|p, slot, *| \times |*, slot, w|}) \quad (1)$$

$$sim(slot_1, slot_2) = \frac{\sum_{w \in T(p_1,s) \cap T(p_2,s)} mi(p_1,s,w) + mi(p_2,s,w)}{\sum_{w \in T(p_1,s)} mi(p_1,s,w) + \sum_{w \in T(p_2,s)} mi(p_2,s,w)} \quad (2)$$

$$S(p_1, p_2) = \sqrt{sim(SlotX_1, SlotX_2) \times sim(SlotY_1, SlotY_2)} \quad (3)$$

3. Searching for the most similar paths

- Computing the similarity of the input path with all of the extracted paths is intractable.
- DIRT makes the search space smaller using some heuristics.
- Then computes the similarity between the input path and remaining paths.

Approach

BERT-Informed Rule Discovery (BIRD)

- We extend DIRT by introducing two path similarity measures.
- They are computed using the contextualized embeddings provided by BERT.

Unweighted BIRD

- Each sentence of a path is fed to a BERT model in order to obtain contextual embeddings of slot-fillers (e_{SlotX} and e_{SlotY}).

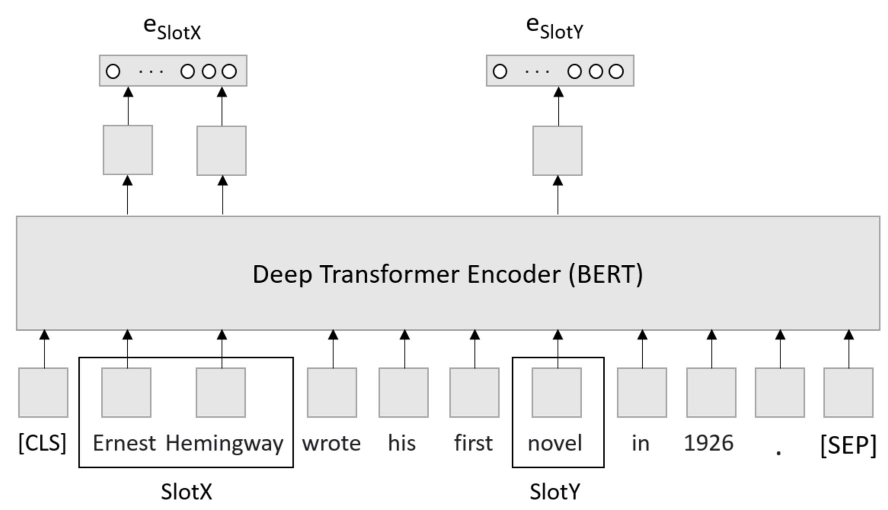

Figure 2: An example sentence is fed to a BERT model. Slot-fillers can be single-word or multi-word noun phrases. In the example sentence, SlotX contains two words (`Ernest Hemingway`) and SlotY contains a single word (`novel`).

- Then the similarity between two paths is computed:

$$\mathrm{E}_s = \frac{1}{n} \sum_i e_{s_i} \quad (4)$$

$$sim(slot_1, slot_2) = \textit{cosine similarity}(E_{s_1}, E_{s_2}) \quad (5)$$

$$sim(p_1, p_2) = \frac{sim(SlotX_1, SlotX_2) + sim(SlotY_1, SlotY_2)}{2} \quad (6)$$

Weighted BIRD

- Some slot-fillers may be more informative than others.
- Mutual information provides additional insight into the importance of slot-fillers.
- Replace Equation 4 with Equation 7.

$$E_s = \sum_i mi(p, s, w_i) \times e_{s_i} \quad (7)$$

Experimental Setting

Q#	Question	English Representation of Path	Path
Q1	Who is the author of the book, "The Iron Lady"?	X is author of Y	`nsubj`←author→`nmod_of`
Q2	What was the price of the Nobel Peace Prize in 1989?	X is price of Y	`nsubj`←price→`nmod_of`
Q3	What does the Peugeot company manufacture?	X manufactures Y	`nsubj`←manufacture→`dobj`
Q4	How much did Mercury spend on advertising in 1993?	X spends Y	`nsubj`←spend→`dobj`
		spends X on Y	`dobj`←spend→`nmod_on`
Q5	What is the name of the director of Apricot Computer?	X is director of Y	`nsubj`←director→`nmod_of`
Q6	Why did David Koresh ask the FBI for a word processor?	X asks Y	`nsubj`←ask→`dobj`
		asks X for Y	`dobj`←ask→`nmod_for`
		X asks for Y	`nsubj`←ask→`nmod_for`

Table 1: The first six TREC-8 questions used for evaluation. Each question is accompanied by its corresponding syntactic path that becomes the seed path for that question.

Experimental Results

Q#	Path	Unweighted BIRD (out of 40)	Weighted BIRD (out of 40)	DIRT (out of 40)
Q1	X is author of Y	23	**24**	20
Q2	X is price of Y	18	**21**	9
Q3	X manufactures Y	32	**33**	30
Q4.1	X spends Y	7	**9**	**9**
Q4.2	spends X on Y	13	12	**19**
Q4.2e	spends X on Y	**25**	24	**25**
Q5	X is director of Y	**17**	**17**	16
Q6.1	X asks Y	15	**16**	8
Q6.2	asks X for Y	**14**	13	4
Q6.3	X asks for Y	21	**25**	15

Table 2: The number of correct paths found by the two BIRD variants compared against the original DIRT algorithm (our implementation).

Acknowledgements

This work was partially supported by NSF grant #2006583, and by DARPA under the World Modelers program, grant number W911NF1810014. Mihai Surdeanu declares a financial interest in lum.ai. This interest has been properly disclosed to the University of Arizona Institutional Review Committee and is managed in accordance with its conflict of interest policies.

References

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3):146–162.

Lin, D. and Pantel, P. (2001). Dirt@ sbt@ discovery of inference rules from text. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 323–328.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.