Why we still need Grammars for NLP

Mark Steedman

(*with* Javad Hosseini (ATI, Edinburgh))

July 14th 2020

Outline

- I: The Grammar vs. the Model in NLP.
- II: Two Approaches to mining Meaning Postulates.
- III: Machine Reading: Results So Far
- IV: Large Language Models: A Comparison
- V: Conclusion

I: The Grammar vs. The Model in NLP

- Any theory of natural language (NL) comprises two modules, the grammar and the model:
  - The grammar defines the semantics;
  - The language model resolves the ambiguity in NL as to which meaning is in play.
- There is always a question as to which of the two is responsible for any phenomenon under discussion.

Semantic Parsing

- In the case of semantic parsing, it has recently become clear that sequence-to-sequence transducers, in which the model bears the whole of the responsibility for mapping strings to trees, perform as well if not better than rule-based parsers for the same amount of training data.
- This fact actually reflects the weakness of parsing models of any kind based on only 1M words of WSJ training data.
- State-of-the-art semantic parsers overcome this limitation by using the labeled data to "fine tune" huge unsupervised language models based on word-embeddings trained on unimaginably vast amounts of unlabeled training data.
- However, notice that we still need structured labels on the semantic side.
- Without a grammar for those structures, it is hard for Seq2Seq to generalize to unseen examples.

Semantic Parsing

- The triumph of the model in semantic parsing raises the question of whether the embeddings that are so effective in disambiguating words for that purpose might also represent word-meaning.
- Linear-algebraic operations such as vector addition and multiplication might then provide compositionality in semantic representation.
- Specifically, it has been suggested that a sequence-to-sequence model based on RoBERTa (Liu *et al.*, 2019) contextualized embeddings embodies a latent model of entailment relations or meaning postulates between predicates, such as that *company1 buying company2* entails *company1 owning company2* (Forbes *et al.*, 2019).

II: Two Approaches to Mining Meaning Postulates

- This talk will compare two approaches to mining entailment relations or meaning postulates:
  - Our own unsupervised approach (Hosseini et al., 2018; 2019, Hosseini, 2020), based on the distributional inclusion hypothesis (DIH, Geffet and Dagan, 2005) over predicates grounded in vectors of named entity argument tuples collected by machine-reading unlabeled text.
  - The supervised language model-based approach of Schmitt and Schütze (2021). trained on corpora of entailments/non-entailments.

III: Our Approach: the Distributional Inclusion Hypothesis

- Use semantic parsers to Machine-Read multiple relations over Named Entities in unlabeled news text.
- Capture relations of entailment and paraphrase over relations between NEs of the same types (Lewis and Steedman, 2013a,b, 2014; Lewis, 2015).
  - If you read somewhere that a a company—say, Google—bought another company—say, YouTube—than you are highly likely to also read somewhere that that company owns that other company—
  - —but not the other way round.
- Redefine the parser semantics in terms of entailments and paraphrases, and reparse and index the entire text for QA.

Local Entailment Probabilities

- First, the typed named-entity technique is applied to (errorfully) estimate local probabilities of entailments:

a. $p(buy\,x\,y \Rightarrow acquire\,x\,y) = 0.9$
b. $p(acquire\,x\,y \Rightarrow own\,x\,y) = 0.8$
c. $p(acquisition\,(of\,x)\,(by\,y) \Rightarrow own\,x\,y) = 0.8$
d. $p(acquire\,x\,y \Rightarrow acquisition\,(of\,x)\,(by\,y)) = 0.7$
e. $p(acquisition\,(of\,x)\,(by\,y) \Rightarrow acquire\,x\,y) = 0.7$
f. $p(buy\,x\,y \Rightarrow own\,x\,y) = \mathbf{0.4}$
g. $p(buy\,x\,y \Rightarrow buyer\,(of\,x)\,y) = 0.7$
h. $p(buyer\,(of\,x)\,y \Rightarrow buy\,x\,y) = 0.7$
i. $p(inherit\,x\,y \Rightarrow own\,x\,y) = 0.7$
(etc.)

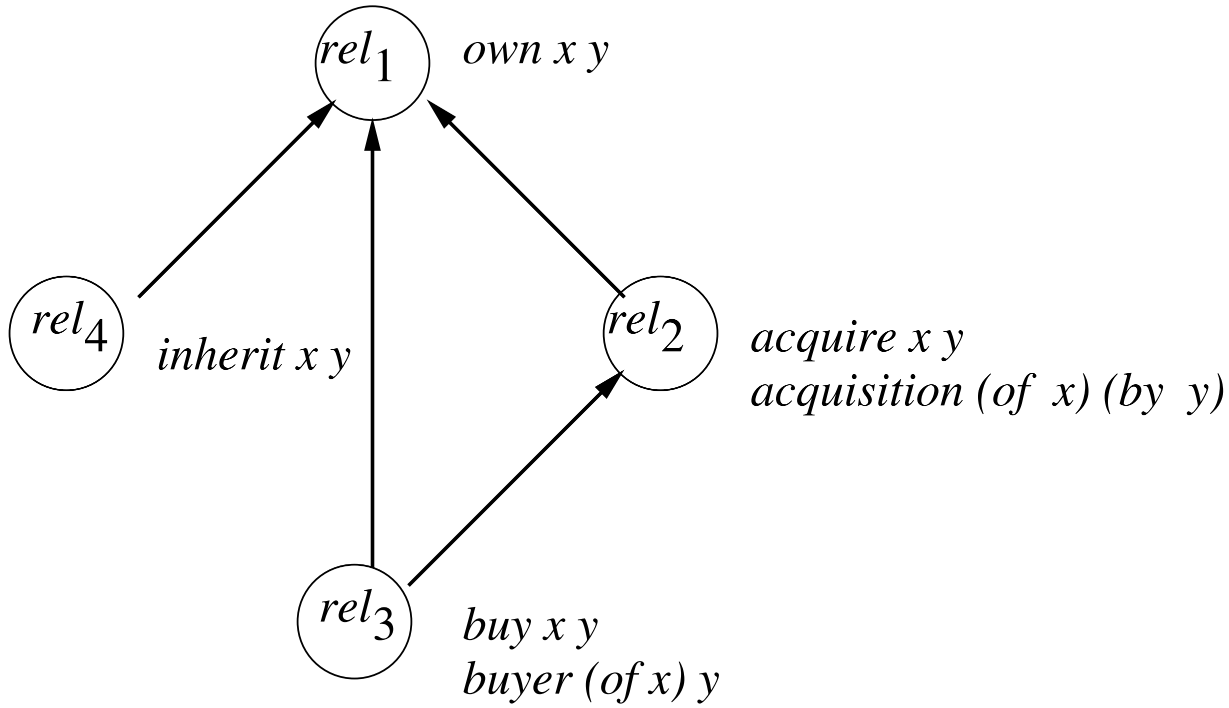
Global Entailments

- The local entailment probabilities are used to construct an entailment graph, with the global constraint that the graph should be closed under transitivity (Berant *et al.*, 2015).

- Thus, local entailment (f) is supported by transitivity despite low observed frequency, while unsupported spurious low frequency local entailments can be excluded.

- Cliques within the entailment graphs can be collapsed to a single paraphase cluster relation identifier.

Entailment graph



- A simplified entailment graph for relations between people and property.

Lexicon

- The new semantics obtained from the entailment graph replaces form-dependent relations like *acquire* with paraphrase cluster identifiers like rel_2

own	:=	$(S\backslash NP)/NP$	$: \lambda x \lambda y.rel_1 x y$
inherit	:=	$(S\backslash NP)/NP$	$: \lambda x \lambda y.rel_4 x y$
acquire	:=	$(S\backslash NP)/NP$	$: \lambda x \lambda y.rel_2 x y$
buy	:=	$(S\backslash NP)/NP$	$: \lambda x \lambda y.rel_3 x y$
buyer of	:=	N/PP_{of}	$: \lambda x \lambda y.rel_3 x y$
etc.			

- These logical forms support correct inference under negation, such as that *Verizon bought Yahoo* entails *Verizon acquired Yahoo* and *Verizon doesn't own Yahoo* entails *Verizon didn't buy Yahoo*

Applications

1. Question Answering.

2. Reranking machine Summarization.

3. Building Knowledge Graphs from text.

Progress So Far

- We have trained an entailment graph on the NewsSpike corpus
  - 0.5M multiply-sourced news articles over 2 months, 20M sentences.
  - 29M binary relation tokens extracted using the CCG parser.
- We have built a working typed global entailment graph, collapsing paraphrase cliques
  - 101K relation types
  - 346 local typed entailment subgraphs
  - 23 subgraphs with more than 1K nodes e.g. Person×Location, Location×Thing, Org×Org, etc.
  - 7 subgraphs with more than 10K nodes
- We redefined the semantics and have built a scalable knowledge graph

Idioms, Metaphors, and Presuppositions

- Idioms are found just like any other typed entailment:
  - $keep\_tabs\_on\,(\#government\_agency, \#thing) \models' s\_surveillance\_of\,(\#government\_agenc$
- So are metaphors:
  - $take\_shot\_at\,(\#person, \#person) \models slam\,(\#person, \#person)$
- Likewise light verbs, particle verbs, etc.:
  - $call\_up(\#person, \#thing) \models work\_with\,(\#person, \#thing)$
- Presuppositions are relations entailed by another relation and its negation:
  - $manage\_to(\#person, \#event) \models try\_to\,(\#person, \#event)$
  - $\neg manage\_to(\#person, \#event) \models try\_to\,(\#person, \#event)$

Intrinsic Evaluation Datasets

- We evaluate on Levy/Holt's (Levy and Dagan, 2016) crowd-annotated entailment dataset
  - Improved by (Holt, 2018), adding inverse pairs and redoing the crowd annotation, which was errorful.
  - 18407 entailment pairs (3916 positively entailing, 14491 nonentailing).
- We also evaluate on Berant's dataset (Berant *et al.*, 2011), obtained by hand-building a gold-standard entailment graph for all parsed relations in their dataset for 10 frequent n-tuples of types, then comparing the extracted graph with this gold-standard.
  - 39012 entailment pairs (3472 positively entailing, 35585 nonentailing).

Refining the Entailment Graph

- Major problem with existing entailment graph learners:
  - Many correct edges are missing because of data sparsity
- Berant *et al.* (2011) used Integer Linear Programming (ILP) to learn entailment graphs, using transitivity closure on the entailments as the objective function: $P \rightarrow Q$ and $Q \rightarrow R$ implies that $P \rightarrow R$.

ILP does not scale to graphs with more than 100 nodes.

- Berant *et al.* (2015) propose an approximation, removing entailment links to make the graph "Forest-Reducible".

FRG loses many valid entailments.

Global Learning of Typed Entailment Graphs

- Instead we propose a scalable method that does not depend on transitivity, but instead uses two global soft constraints.
  - Our method scales to more than 100K nodes.

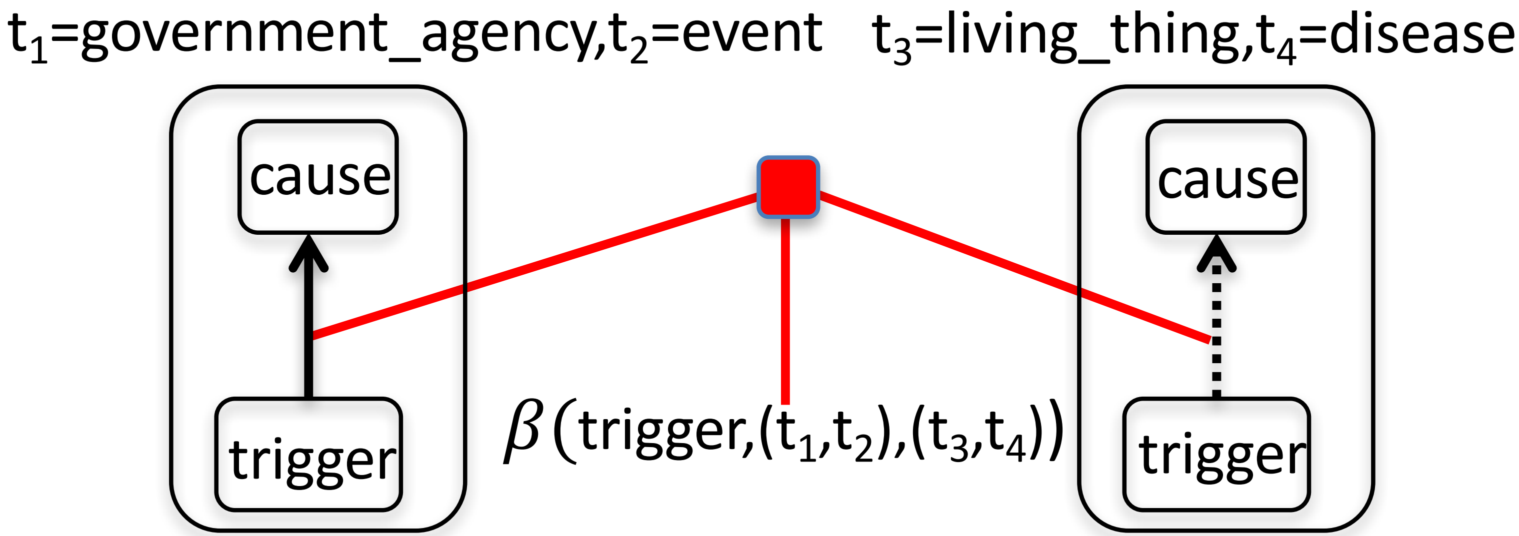
Global Soft Constraint 1: Cross Graph Transfer

- It is standard to learn a separate typed entailment graph for each (plausible) type-pair Berant *et al.* (2011, 2012); Lewis and Steedman (2013a,b); Berant *et al.* (2015).

- However, many entailment relations for which we have direct evidence only in a *few* subgraphs may apply over *many* others.

- This is a form of Domain Tramsfer.

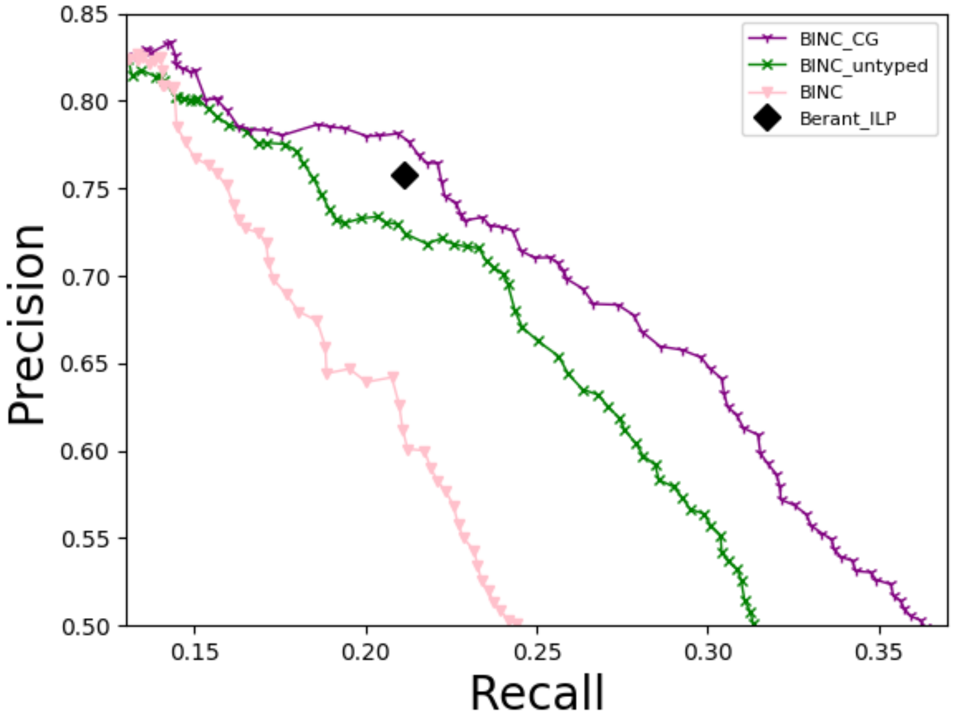
Global Soft Constraint 1: Cross Graph



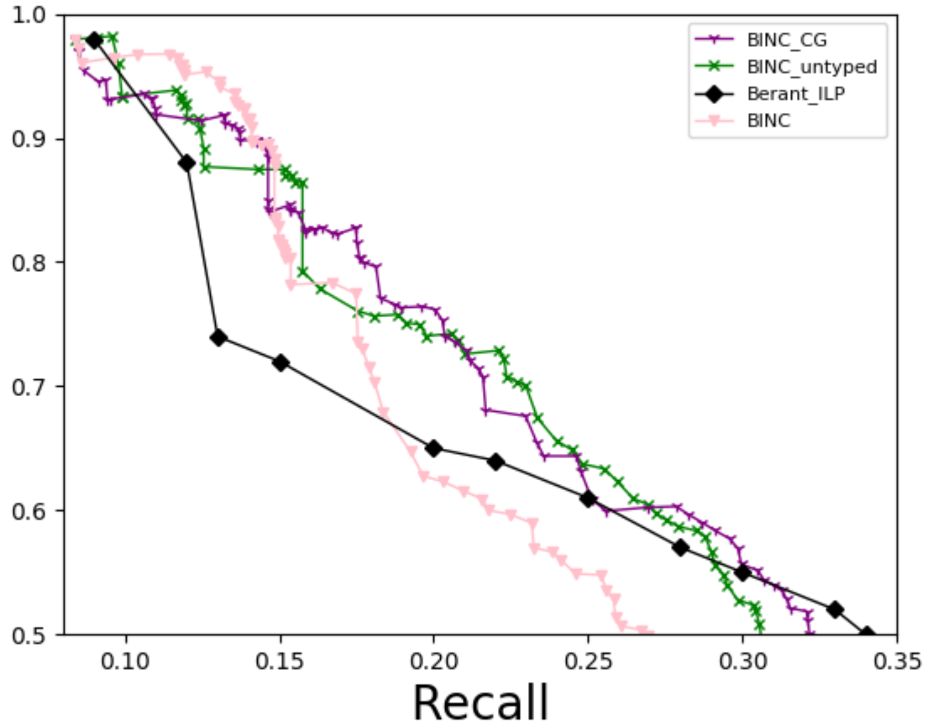
- $0 \leq \beta(.) \leq 1$ determines how much different graphs are related and will be learned jointly.

Adding Cross-Graph Transfer Soft Constraints

Levy/Holt's dataset

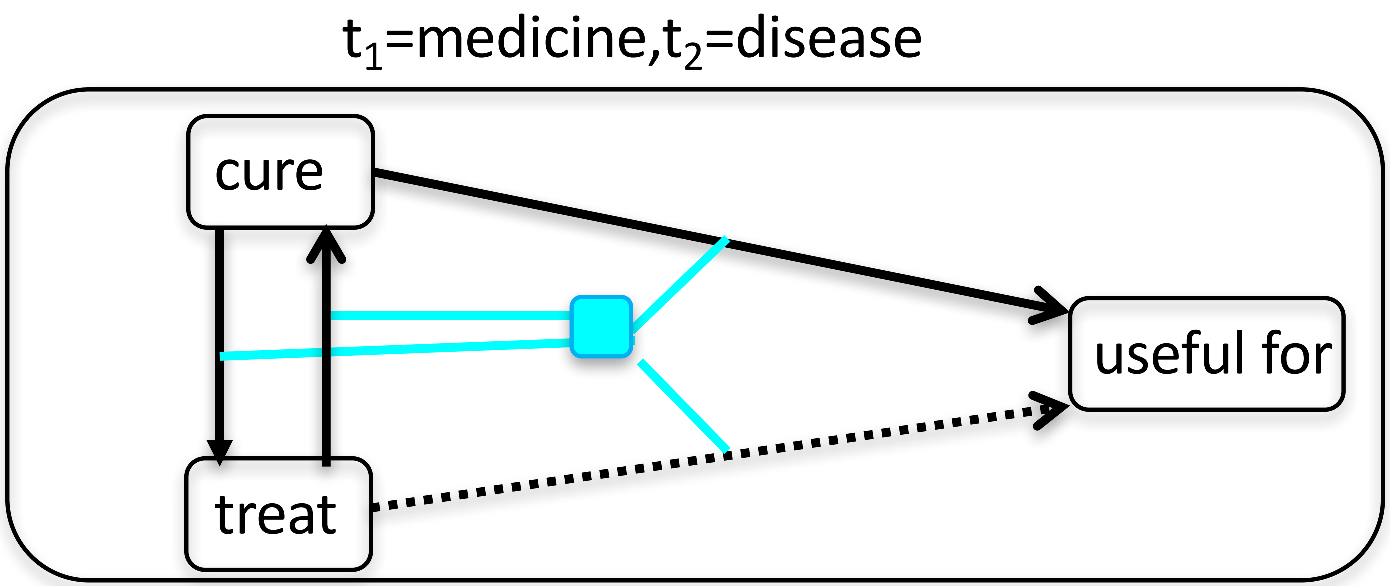


Berant's dataset



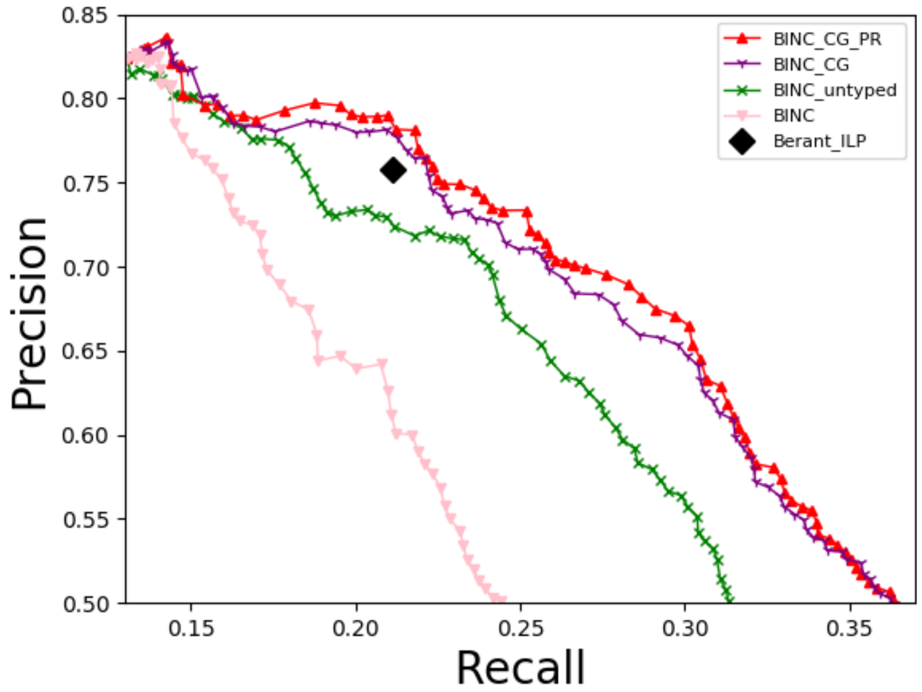
Global Soft Constraint 2: Paraphrase Resolution

- We encourage paraphrase predicates (where $i \rightarrow j$ and $j \rightarrow i$) to have the same patterns of entailment
  - i.e. to entail and be entailed by the same predicates

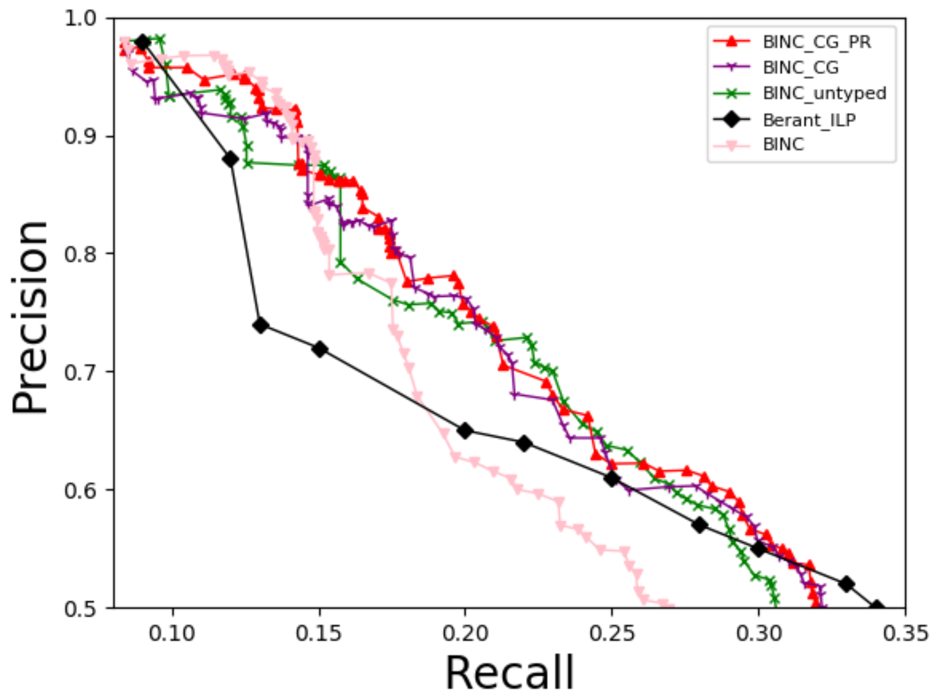


Adding Paraphrase Resolution Soft Constraints

Levy/Holt's dataset



Berant's dataset



Results for Various Similarity Measures

- Area under precision-recall curve (precision $> .5$) for different variants of distributional similarities
  - Boldfaced results are statistically significant

	local	untyped	CG	CG_PR
	Levy/Holt's dataset			
BInc	.076	.127	.162	**.165**
Lin	.074	.120	**.151**	**.149**
Weed	.073	.115	**.149**	**.147**
	Berant's dataset			
BInc	.138	.167	**.177**	**.179**
Lin	.147	.158	.186	**.189**
Weed	.146	.154	.184	**.187**

Example Subgraph after CG and PR

Premise	Entails	Consequents
location suffers from *thing*	$\rightarrow$	*thing* killing in *location* *location* has *thing* *location* 's price for *thing* *location* suffers *thing* *location* diagnosed with *thing* destroyed during *thing* in *location* *thing* affects *location* *thing* 's image in *location* *location* recovers *thing* *location* 's *thing* *location* experiences *thing* took across *location* in *thing*

Test: Africa suffers from droughts $\rightarrow$ Africa experienced a drought Correct

Error Analysis

Error type	Example
	False Positive
High correlation (57%)	Microsoft **released** Internet Explorer $\rightarrow$ Internet Explorer **was developed by** Microsoft
Relation normalization (31%)	The pain **may be** relieved by aspirin $\rightarrow$ The pain can be treated with aspirin
Lemma baseline & parsing (12%)	President Kennedy **came** to Texas $\rightarrow$ President Kennedy **came** from Texas
	False Negative
Sparsity (93%)	Cape town **lies at the foot of** mountains $\rightarrow$ Cape town is located near mountains
Wrong label & parsing (7%)	Horses are imported from Australia $\rightarrow$ Horses are native to Australia

Extrinsic Evaluation

- We have carried out a limited extrinsic evaluation on an answer selection task on the NewsQA test set of text-questions (Trischler *et al.*, 2017), achieving a 1-2% increase in performance over a baseline inverse sentence frequency (ISF) measure (cf. Narayan *et al.*, 2018).

	ACC	MRR	MAP
ISF	.3618	.4899	.4857
ISF+ENT	**.3761**	**.5006**	**.4963**

Table 1: Answer selection on NewsQA

- NewsQA example:

  Question: Who praised Mitt Romney's credentials?
  Selected sentence: The board hailed Romney for his solid credentials

Do Embeddings Help?

- Rather than guessing entailment relations based on directional similarity of vectors of named-entity pairs, our colleagues frequently ask us, why not try the "alternative approach", representing relations as embeddings, and applying a directional distributional inclusion similarity measure
- We keep trying this. It hasn't worked yet.
- However, Hosseini *et al.* (2019) show that embeddings-based methods for link-prediction in existing knowledge graphs (Riedel *et al.*, 2013) can be used to replace the PMI measure with normalized link prediction scores derived from the extracted triples to improve the local graph before globalization.
- And vice versa—access to the entailment graph improves link-prediction.

Do Embeddings Help?

- Hosseini (2020); Hosseini *et al.* (2021) shows that contextualized embeddings can be applied to the actual context from which each parsed triple has been mined, and used in the same way to build the local entailment graph
- The embeddings seem to embody a latent type-system that in some cases compensates for the weakness of FIGER entity typing in earlier work (Choi *et al.*, 2018)
- Embeddings seem to learn information that is complementary to machine-reading.
- This version of the pipeline has been applied to an order of magnitude more news data (NewsCrawl), improving performance (results below).

IV: The Large Language Model-based Approach

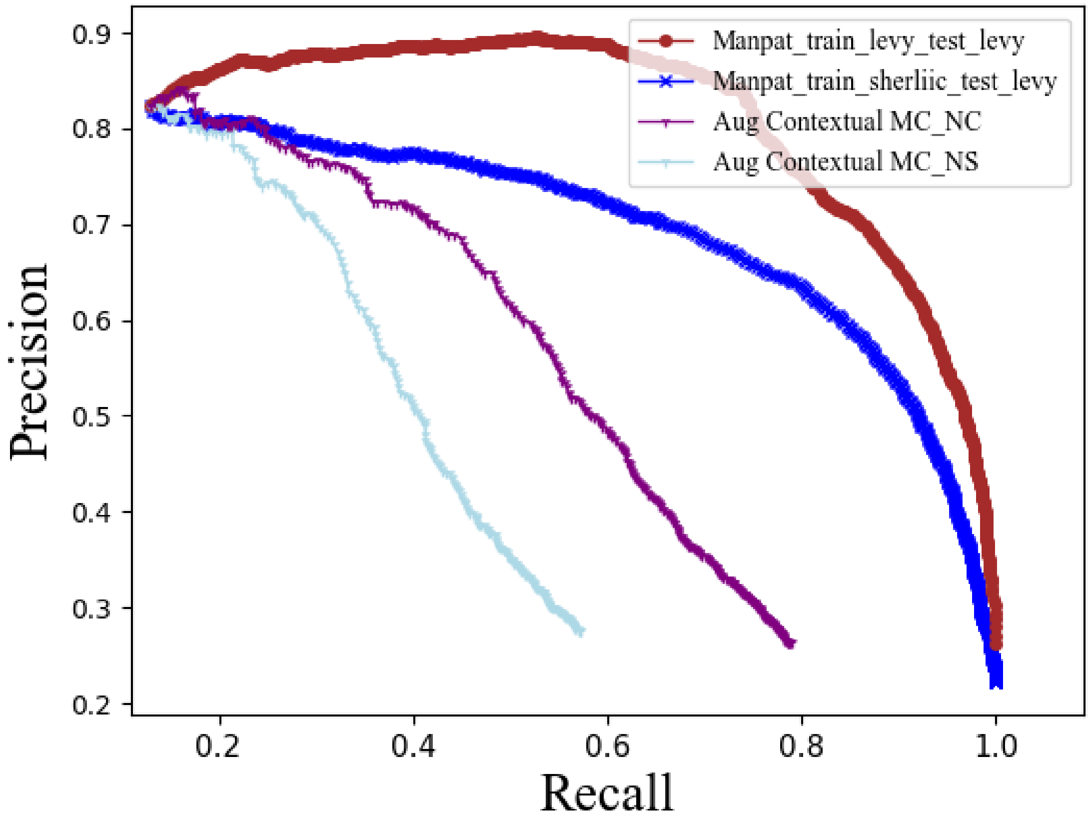
- Schmitt and Schütze (2021) make a direct comparison of their RoBERTa model-based approach with our unsupervised DIH approach.
- Their approach is supervised, training and testing on two entailment-pair datasets: their own small SherLIiC corpus and our own larger Levy-Holt.
- They follow Amrami and Goldberg (2018) in using manually selected entailment Hearst patterns such as "P because Q" to induce directionality of entailment.
- They show the following AUC table (corrected):

Hosseini et al. 2018	28.4
Hosseini et al. 2019	30.6
S&SmanpatRoBERTabase	**76.9**
S&SmanpatRoBERTalarge	**83.9**

- However these numbers do not tell the whole story.

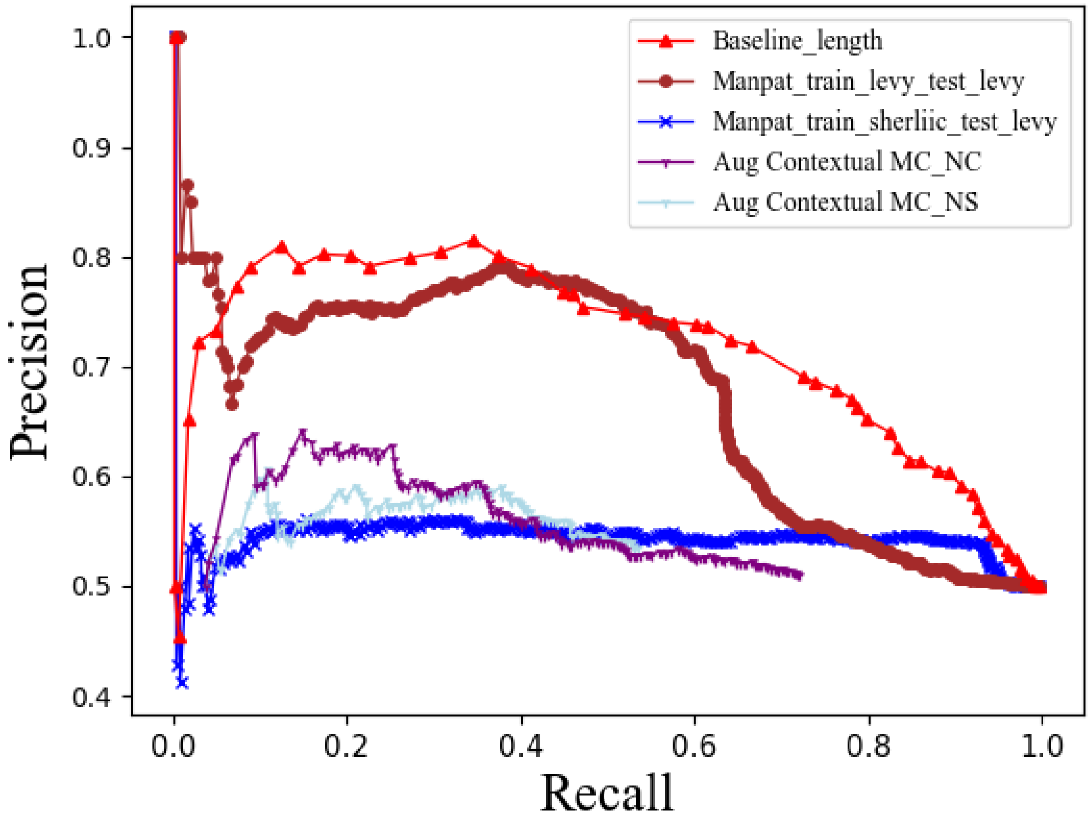
What is the Model Actually Learning?

- Supervised end-to-end machine learning is notorious for picking up any artefacts in the training data that are predictive of the labels.
- Poliak *et al.* (2018) found that NLI programs trained on the hypotheses alone did as well on entailment test sets as those trained on the full entailments.
- What happens if we train on SherLIiC and test on Levy-Holt?
- A lot of their AUC advantage goes away.
- In particular our DIH is better at the High Precision end, while S&S is better at the High Recall end.


Manpat_train_levy_test_levy
Manpat_train_sherliic_test_levy
Aug Contextual MC_NC
Aug Contextual MC_NS
Precision
Recall
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.2
0.4
0.6
0.8
1.0

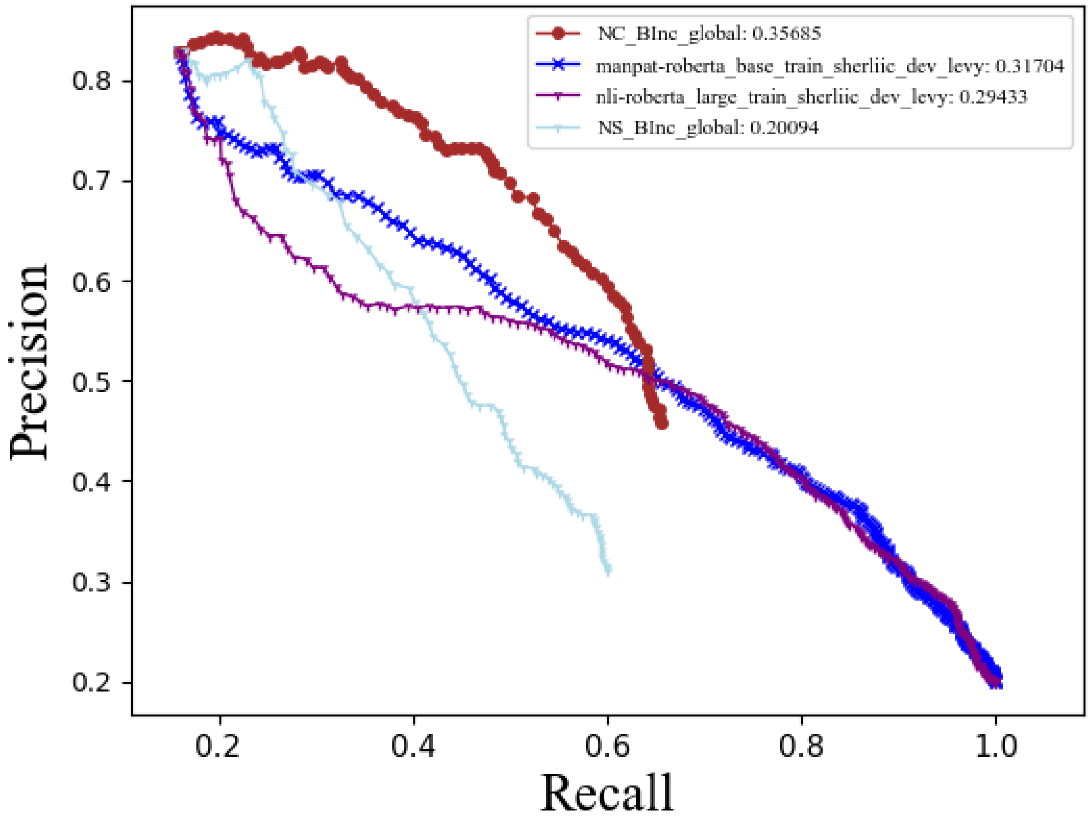
What is the Model Actually Learning?

- But is S&S actually learning entailment at all?
- As in any handbuilt entailment dataset, there are artefacts in Levy-Holt—that is why we never train on it.
- In particular, there is a length bias for true positives.
- The length heuristic alone beats S&S trained on Levy-Holt on standard Levy-Holt test, suggesting that this is what is being learned.
- If we train S&S on SherLIic and test only on the subset of Levy-Holt test set that is actually directional, where $P \models Q$ but $Q \not\models P$, performance drops to near chance.
- Performance of DIH is actually better in the range 0-0.5 recall.


Baseline_length
Manpat_train_levy_test_levy
Manpat_train_sherliic_test_levy
Aug Contextual MC_NC
Aug Contextual MC_NS
Precision
Recall
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.0
0.2
0.4
0.6
0.8
1.0

Conclusion

- The above results are unsurprising: Embeddings are essentially Associative, rather than Semantic, in nature.
- Contextualized embeddings like RoBERTa are, as we see in the case of semantic parsing, extremely effective at disambiguating words and other categories on the basis of similarity of their current context to contexts seen in the vast unlabeled training data.
- The largest of these models have up to a trillion parameters, and can memorize the training data (Zhang *et al.* 2021).
- Fine-tuning with small amounts of labeled data seems to tell them which region of memory to access to simulate your task.
- I see no evidence so far that this ability extends to representing word-meaning in the sense of supporting inference.


NC_BInc_global: 0.35685
manpat-roberta_base_train_sherliic_dev_levy: 0.31704
nli-roberta_large_train_sherliic_dev_levy: 0.29433
NS_BInc_global: 0.20094
Precision
0.8
0.7
0.6
0.5
0.4
0.3
0.2
Recall
0.2
0.4
0.6
0.8
1.0

Thanks!

• To: ERC Advanced Fellowship SEMANTAX; ARC Discovery grant DP160102156; Huawei/Edinburgh Research; Google Faculty Award; Bloomberg L.P. Gift Award.

References

Amrami, Asaf and Goldberg, Yoav, 2018. "Word Sense Induction with Neural biLM and Symmetric Patterns." In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 4860–4867.

Berant, Jonathan, Alon, Noga, Dagan, Ido, and Goldberger, Jacob, 2015. "Efficient Global Learning of Entailment Graphs." *Computational Linguistics* 42:221–263.

Berant, Jonathan, Dagan, Ido, Adler, Meni, and Goldberger, Jacob, 2012. "Efficient Tree-Based Approximation for Entailment Graph Learning." In *Proceedings of the 50th Annual Meeting of the Association for Computational*

Linguistics: Long Papers-Volume 1. Association for Computational Linguistics, 117–125.

Berant, Jonathan, Goldberger, Jacob, and Dagan, Ido, 2011. "Global Learning of Typed Entailment Rules." In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*. Portland, OR, 610–619.

Choi, Eunsol, Levy, Omer, Choi, Yejin, and Zettlemoyer, Luke, 2018. "Ultra-Fine Entity Typing." In *Proceedings of the 56th Annual Conference of the Association for Computational Linguistics*. ACL, 87–96.

Forbes, Maxwell, Holtzman, Ari, and Choi, Yejin, 2019. "Do Neural Language Representations Learn Physical Commonsense?" In *Proceedings of the 41st Annual Meeting of the Cognitive Sciience Society*. 1753–1759.

Geffet, Maayan and Dagan, Ido, 2005. "The Distributional Inclusion Hypotheses and Lexical Entailment." In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*. ACL, 107–114.

Holt, Xavier, 2018. *Probabilistic models of relational implication*. Master's thesis, Macquarie University, Sydney.

Hosseini, Javad, 2020. *Learning Typed Entailment Graphs from Text*. Ph.D. thesis, University of Edinburgh.

Hosseini, Javad, Chambers, Nathaniel, Reddy, Siva, Ricketts-Holt, Xavier, Cohen, Shay, Johnson, Mark, and Steedman, Mark, 2018. "Learning Typed Entailment Graphs with Global Soft Constraints." *Transactions of the Association for Computational Linguistics* 6:703–718.

Hosseini, Javad, Cohen, Shay, Johnson, Mark, and Steedman, Mark, 2019. "Duality of Link Prediction and Entailment Graph Induction." In *Proceedings* *of the 57th Annual Conference of the Association for Computational Linguistics (long papers)*. ACL, 4736–4746.

Hosseini, Javad, Cohen, Shay, Johnson, Mark, and Steedman, Mark, 2021. "Contextual Link Prediction to Learn Relational Entailment Graphs." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. in preparation.

Levy, Omer and Dagan, Ido, 2016. "Annotating Relation Inference in Context via Question Answering." In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. ACL, 249–255.

Lewis, Mike, 2015. *Combined Distributional and Logical Semantics*. Ph.D. thesis, University of Edinburgh.

Lewis, Mike and Steedman, Mark, 2013a. "Combined Distributional and Logical Semantics." *Transactions of the Association for Computational Linguistics* 1:179–192.

Lewis, Mike and Steedman, Mark, 2013b. "Unsupervised Induction of Cross-Lingual Semantic Relations." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. ACL, 681–692.

Lewis, Mike and Steedman, Mark, 2014. "Combining Formal and Distributional Models of Temporal and Intensional Semantics." In *Proceedings of the ACL Workshop on Semantic Parsing*. Baltimore, MD: ACL, 28–32. Google Exceptional Submission Award.

Liu, Yinhan, Ott, Myle, Goyal, Naman, Du, Jingfei, Joshi, Mandar, Chen, Danqi, Levy, Omer, Lewis, Mike, Zettlemoyer, Luke, and Stoyanov, Veselin,

2019. “RoBERTa: A Robustly Optimized BERT Pretraining Approach.” *arXiv preprint arXiv:1907.11692* .

Narayan, Shashi, Cardenas, Ronald, Papasarantopoulos, Nikos, Cohen, Shay, Lapata, Mirella, Yu, Jiangsheng, and Chang, Yi, 2018. “Document Modeling with External Attention for Sentence Extraction.” In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL, 2020–2030.

Poliak, Adam, Naradowsky, Jason, Haldar, Aparajita, Rudinger, Rachel, and Van Durme, Benjamin, 2018. “Hypothesis-only Baselines in Natural Language Inference.” *Proceedings of the Seventh NAACL-HLT Joint Conference on Lexical and Computational Semantics* :180–181.

Riedel, Sebastian, Yao, Limin, McCallum, Andrew, and Marlin, Benjamin, 2013. “Relation Extraction with Matrix Factorization and Universal Schemas.” In

Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Atlanta: ACL, 74–84.

Schmitt, Martin and Schütze, Hinrich, 2021. "Language Models for Lexical Inference in Context." In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 1267–1280.

Steedman, Mark, 2000. *The Syntactic Process*. Cambridge, MA: MIT Press.

Trischler, Adam, Wang, Tong, Yuan, Xingdi, Harris, Justin, Sordoni, Alessandro, Bachman, Philip, and Suleman, Kaheer, 2017. "NewsQA: A Machine Comprehension Dataset." In *Proceedings of the 2nd Workshop on Representation Learning for NLP*. ACL, 191–200.

Zhang, Chiyuan, Bengio, Samy, Hardt, Moritz, Recht, Benjamin, and Vinyals, Oriol, 2021. "Understanding Deep Learning (Still) Requires Rethinking Generalization." *Communications of the ACM* 64:107–115.

Zhang, Congle and Weld, Daniel, 2013. "Harvesting Parallel News Streams to Generate Paraphrases of Event Relations." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Seattle: ACL, 1776–1786.