Linguistically informed tasks for evaluating structure encoded by sentence representations

Najoung Kim [1] Benjamin Van Durme [2] Ellie Pavlick [3] Paul Smolensky [1,4]

[1]Department of Cognitive Science, Johns Hopkins University [2]Center for Language and Speech Processing, Johns Hopkins University [3]Department of Computer Science, Brown University [4]Microsoft Research AI

What do we want from modeling language?

General consensus: Doing well on one downstream task is not equivalent to knowing language.

- Do we want models that have human-like linguistic capacity?
- Or do we want models that are optimized for a particular problem?

Depends on the circumstance, but...

- ...we at least agree that that human linguistic ability is *general-purpose*.
- This is probably useful, so assuming that the goal of modeling language is to approximate human linguistic capacity (i.e., *knowing* language):

Doing well on a specific downstream task ≠ knowing language
Doing well on multiple downstream tasks ?= knowing language
☞ Doing well on linguistic tasks that humans can easily solve ≈ knowing language (e.g., Mikolov et al. (2013)'s analogy task)

But what tasks?
⟹ Tasks that probe for how well function words & structure are understood; i.e.,

What the models should know to be better than bag-of-words!

Preliminary set of evaluation tasks

- **Structured mutation** of an existing dataset focusing on function words & structure
- **Easy for humans** (Try these yourselves! ☺)

(1) Locative preposition swap

(Premise) The tomb of Mohammed Ali sits under the colonnade.
(Original Hypothesis) The colonnade stands above a famous tomb. — entailment

(Premise) The tomb of Mohammed Ali sits under the colonnade.
(Modified Hypothesis) The colonnade stands beneath a famous tomb. — contradiction

(2) Lexical/explicit negation

I was lying on a dirty bed (P) - I was on a clean bed (H) — contradiction
I was lying on a dirty bed (P) - I was not on a clean bed (¬H) — entailment
I was lying on a dirty bed (P) - I was on a dirty bed (¬H) — entailment
I was lying on a dirty bed (P) - I was not on a dirty bed (¬¬H) — contradiction

(3) *wh*-word identification

A man whxxxx knows how to fish can feed himself for his entire life. — A:'who'
a woman sits next to a harbor whxxxx boats are docked — A:'where'

(4) Definite-indefinite articles

He was also named the Canadian Player of the Month in July — ✓
He was also named a Canadian Player of a Month in July — X

(5) Possessor-possessee distinction

Each one planting itself in the sides of xxxxx 's xxxxx . {Stark's neck (✓), neck's Stark (X)}

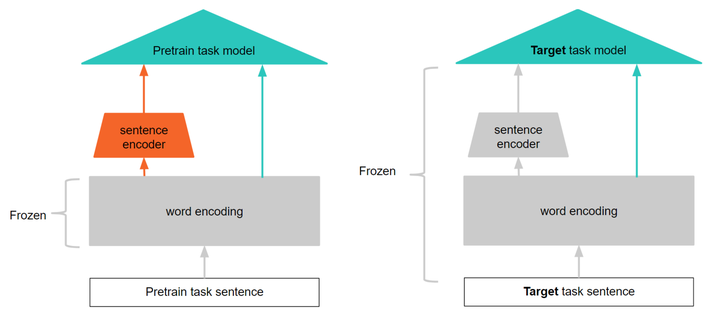
(6) End-of-sentence identification

so i turned some good contracts down i let it all slide until the middle of january
when i called them up and said well where the contracts — down | i

Evaluation setup

Goal: Compare what different tasks teach models about function words & structure.

1. Train a BiLSTM sentence encoder with a *pretraining* task (hyperparameters are fixed).
2. Freeze the encoder weights.
3. Train an MLP on top of the frozen encoder for the target (evaluation) tasks.
4. Compare the effects of various pretraining tasks [2, 3, 4, 5, 6] on target tasks.



Result 1: Tasks differentiate between pretraining tasks

MNLI (inference) helps with negation and definiteness, and CCG (almost-syntax) helps with wh-words and sentence structure. MT helps with locative prepositions.

Encoder training tasks → Evaluation tasks ↓	Random init.	MNLI	MT (En-De)	DisSent	CCG	MSCOCO (Grounded)	LM
Training data size	-	393k	3.4M	151k	38k	118k	4M
Locative preposition swap	56.4	57.4	**58.5**	56.1	54.4	55.2	56.2
Lexical & explicit negation	50.0	**56.3**	53.4	48.7	44.6	47.3	42.1
wh-word identification	79.4	81.0	80.5	80.7	**86.5**	78.1	77.8
Definite-indefinite articles	89.8	**92.8**	92.1	90.4	92.4	91.2	90.9
Possessor-possessee distinction	98.2	**98.4**	**98.4**	98.2	97.7	98.2	98.2
EOS identification	12.0	13.5	18.6	15.1	**18.7**	11.3	10.6

Table: For locative preposition & negation tasks, we report accuracy given that the model answered the unmutated question correctly. For others, we report raw accuracy.

Result 2: Doing well on downstream tasks ≠ Doing well on evaluation

Pretraining on CCG performs best on most downstream tasks we tested. However, this does not necessarily mean better performance on the evaluation task set. (Compare with Table 1)

Encoder training tasks → Evaluation tasks ↓	Random init.	MNLI	MT (En-De)	DisSent	CCG	MSCOCO (Grounded)	LM
SRL (CoNLL 2005)	80.1	86.8	87.7	86.6	**88.0**	86.6	86.3
SRL (CoNLL 2012)	78.7	83.1	87.0	82.3	**87.8**	79.2	78.7
Dependency (UD)	90.3	91.9	91.6	92.2	**93.6**	86.7	90.4
Constituency (Ontonotes)	77.9	78.5	80.3	78.9	**81.9**	73.1	78.4
SPR2 (White et al. 2016)	81.8	**82.6**	**82.6**	82.5	82.5	82.2	82.2

TL;DR

Contributions

- Proposed a set of tasks that are easily solvable by humans to evaluate the quality of function words & structure encoded by sentence representations.
- Compared the effect of various pretraining tasks (MNLI, MT, DisSent, CCG, MSCOCO, LM).

Main findings

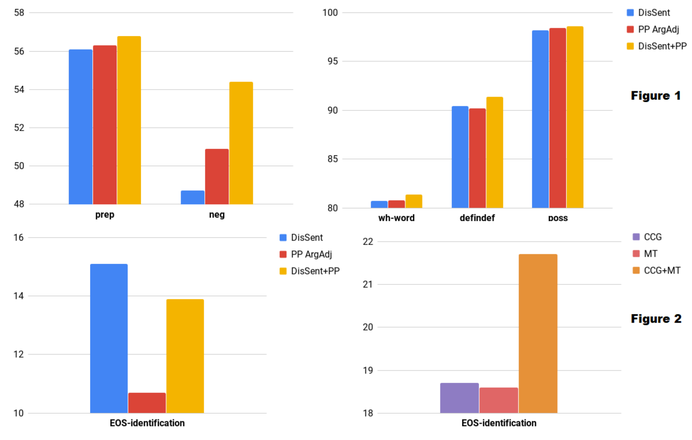
- Our evaluation tasks differentiate between pretraining tasks, and provides interpretable results (e.g., pretraining on MNLI helps understand negation better).
- Multitask-learning on a theoretically motivated task (PP arg. vs adj. distinction) helps models do better on evaluation tasks.
- Doing better on downstream tasks (e.g., SRL, dependency labeling) does not transparently translate into doing better on our evaluation tasks.

Multitask Learning (Result 3: Theoretically motivated task helps)

Added task: PP argument/adjunct distinction (theoretically movitated task):
put x *on* y / shape x *into* y **(arg)** walk *without* x / swim *for* x **(adj)**

Fig. 1: Multitask training with DisSent (underperforming pretraining task) and PP argument/adjunct distinction task improves 5 out of the 6 evaluation task performance.

Fig. 2: Gain from multitask training isn't just from seeing more data; additional data might not help if it is not informative for solving the task (left). We see that if the additional training task is informative (CCG+MT), we can improve performance on EOS also.



References

[1] Mikolov et al. (2013). Linguistic regularities in continuous space word representations. *Proc. of NAACL*. 746--751. [2] Bojar et al. (2014) Findings of the 2014 workshop on statistical machine translation. *Proc. of the 9th workshop on statistical MT. 12--58*. [3] Lin et al. (2014). Microsoft coco: Common objects in context. *European Conference on Computer Vision*. 740--755. [4] Nie et al. (2017). DisSent: Sentence representation learning from explicit discourse relations. arXiv preprint arXiv:1710.04334. [5] Bangalore & Joshi (1999). Supertagging: An approach to almost parsing. *Computational linguistics*, 25(2):237--265. [6] Williams et al. (2018). A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. *Proc. of NAACL*, Vol 1, 1112-1122. [7] White et al. (2016). Universal decompositional semantics on universal dependencies. *Proc. of EMNLP*. 1713--1723. [8] Carreras et al. (2005). Introduction to the CoNLL-2005 shared task: Semantic role labeling. *Proc. of 9th conference on computational natural language learning*. 152--164. [9] Pradhan et al. (2012). CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. *Joint Conference on EMNLP and CoNLL-Shared Task*. 1--40. [10] Nivre et al. (2016). Universal Dependencies v1: A Multilingual Treebank Collection. LREC.