Linguistically informed tasks for evaluating structure encoded by sentence representations

Neural network models have become the standard approach for many tasks in Natural Language Processing (LeCun et al. 2015). However, we still struggle to understand what the models are learning about language, especially models that make use of sequential representations involving compositional structures. Recent efforts (Ettinger et al. 2016; Gulordava et al. 2018; Ribeiro et al. 2018) try to look inside the 'black box' via *probing tasks*, where tasks that require a specific subset of linguistic knowledge (e.g., entailment preservation with synonym substitution, (im)plausibility of thematic role inversion, etc.) in order to be solved are presented to a model, and success (or failure) on such tasks guides qualitative evaluation of the learned representations.

We share the belief that this probing task approach holds much promise. Here we propose a preliminary set of tasks that focus on function words, since function words are crucial in mastering compositionality and hierarchical structure, the lack of which many modern models still suffer from (He et al. 2017). Our tasks use principled transformations of an existing natural language inference dataset (MNLI; Williams et al. 2018). These include structure-preserving mutations (noncing, corruption of non-function words) and substitutions triggering entailment changes (e.g., prepositions, antonyms, negations). The following are selected examples from our probing set:

(1) Locative preposition swap (Original Premise) The tomb of Mohammed Ali sits under the colonnade.
(Original Hypothesis) The colonnade stands *above* a famous tomb. **entailment**
(Locative-swapped Hypothesis) The colonnade stands *beside* a famous tomb. **contradiction**

(2) Corruption (Premise) Siittng up at nighht is alawys rahter jummpy, she cotnfessed .
(Hypothesis) She stated, "Sititng up at npight is relxaing." **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 near people shopping **A:'where'**

(4) 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 ($\neg$H) **entailment**
I was lying on a **dirty** bed (P) - I was on a **dirty** bed ($\neg$H) **entailment**
I was lying on a **dirty** bed (P) - I was **not** on a **dirty** bed ($\neg\neg$H) **contradiction**

We designed our probing set including the above tasks based on the following criteria: (1) performance on the task should depend on the mastery of the targeted type of phenomenon or function words, (2) the tasks should be easy for native speakers with no special training, and (3) the set should have good coverage of closed-class items.

Our contributions are threefold. First, we demonstrate how our tasks can qualitatively differentiate sentence representations by comparing sentence encoders *pre*trained on different tasks: MNLI, machine translation (En-De), CCG supertagging (Bangalore and Joshi 1999), DisSent (Nie et al. 2017). Table 1 shows that the model pretrained on CCG supertagging does better on tasks that require syntactic knowledge (corruption, *wh*-word id) and the MNLI model does better on tasks that require a good understanding of semantics of negation. Second, we show that there is room for theoretically-motivated tasks in improving the representations. We test PP argument/adjunct distinction ("PPArg") as a pretraining task, which is a theoretically important distinction but does not have an immediate downstream application. We find that adding this task ("CCG+PPArg") helps with multiple probing tasks (improvements marked with †). Finally, we observe that better downstream task performance (e.g., Semantic Role Labeling) does not transparently translate into doing well on probing tasks (i.e., what a human would easily be able to do), which casts doubt on relying on popular downstream tasks as metrics of 'better' models of language.

Encoder training tasks →	Random	MNLI	MT	DisSent	CCG	PPArg	CCG+PPArg
Probing tasks ↓	init.					(theoretical)	(augmented)
Locative preposition swap	55.7	57.4	**58.5**	55.5	54.4	54.2	55.9†
Corrupt content words	-11.0	-21.3	-12.6	-11.3	-5.1	-14.7	**-4.1**†
Lexical & explicit negation	42.0	**56.3**	53.4	46.6	44.6	52.3	44.5
wh-word identification	79.4	81.0	80.5	80.7	**86.5**	80.2	85.4
PP attachment (Belinkov et al. 2014)	86.1	86.8	87.7	86.6	88.0	87.3	**88.8**†
SRL (CoNLL 2005)	80.1	87.3	89.1	87.8	89.2	84.3	**90.0**†
SRL (CoNLL 2012)	78.7	83.1	87.0	82.3	**87.8**	78.9	87.2

Table 1: The first four rows of the table show performance on tasks in our probing set (only a subset is listed due to page limitations). The last three rows show results on existing tasks. For the corruption test, we report accuracy drop (%p) compared to an unchanged control set (i.e., smaller absolute values are better). Accuracy is reported for all other tasks, except for micro F_1 for SRL.