

# Supplementary Material

**Pranav Goel**  
Computer Science  
University of Maryland  
pgoell1@cs.umd.edu

**Shi Feng**  
Computer Science  
University of Maryland  
shifeng@umiacs.umd.edu

**Jordan Boyd-Graber**  
Computer Science, iSchool,  
UMIACS, and LSC  
University of Maryland  
jbg@umiacs.umd.edu

## A Training Detail

We specify the variation of each pre-trained representation we used:

1. GloVe: 300 dimensional vectors trained on the Common Crawl dataset with representations for 840B tokens.<sup>1</sup>
2. ELMo: The ‘original’ pre-trained ELMo model using a 2-layer LSTM.<sup>2</sup> The word embedding dimension is 1024.
3. BERT: The fully pre-trained BERT-base model (uncased) using a 12-layer bidirectional transformer.<sup>3</sup>

**We take the mean of the multi-layer BERT/ELMo representations.** The probing model (Figure 1) used 500 dimensional hidden layer(s) with ReLU non-linearities. We train with SGD (learning rate = 0.005) with momentum = 0.7 for 25 epochs with a patience of 20. Our models are implemented using the AllenNLP framework (Gardner et al., 2018).

## B The Reversal Trick

Figure 2 visualizes the reversal trick. Passing the reversed embedding ( $word_1 - word_2$  as well as  $word_2 - word_1$ ) at test time helped Yang et al. (2018) achieve better accuracy on the Verb Physics dev set. With our setup being basically the same as their approach, except for their use of poles, our results that the reversal trick was the major reason for their accuracy and not the use of poles (Table A1). We tried using the reversal trick at training time as well which further increased the scores, making the accuracy of the model better than the ‘PCE’ model (§2.3) without the use of poles (Table 1)!

## C Global Ordering over all Words: Example

The process is described in §4.1.2: the trained weight of a linear model corresponding to label 1 or 0 is multiplied with the pre-trained representation of a word to get a ‘score’ corresponding to every word. This score can then be used to form an ordering over all the words in a particular set. Table A2 contains an example of such an ordering obtained using GloVe for the size attribute over all the words in the corresponding Verb Physics dev set using the trained weight corresponding to the label 1.

<sup>1</sup><https://nlp.stanford.edu/projects/glove/>

<sup>2</sup><https://allennlp.org/elmo>

<sup>3</sup><https://github.com/huggingface/pytorch-pretrained-BERT>

	<i>Concatenation</i>			<i>Subtraction</i>		
	No Reversal	Reversal at Test Only	Reversal at Train & Test	No Reversal	Reversal at Test Only	Reversal at Train & Test
<b>GloVe</b>	0.72	0.74	0.77	0.75	0.74	0.76
<b>ELMo</b>	0.73	0.76	0.78	0.77	0.77	0.77
<b>BERT</b>	0.7	0.74	0.75	0.73	0.74	0.75

Table A1: Effect of using the reversal trick (Figure 2): Accuracy increases for all the representations when using reversal at the test time (as noted by (Yang et al., 2018)), and a further increase when used at both training and testing. The effect is more pronounced when the embeddings are combined via concatenation.

scissors < beard < spoon < hair < knife < finger < lip < purse < chin < goose < vial < eye < nose < bow < fist < piece < ash < glass < chair < skirt < grass < picture < head < face < hat < gulp < bag < ear < hand < strap < dress < bottle < torso < elbow < edition < mouth < pocket < arm < shoulder < rope < magazine < tear < seal < hedge < effect < violin < tree < knee < lamp < cup < pedestal < throat < book < coal < object < suit < button < ball < chest < magistrate < newspaper < fox < ice < candidate < harlot < basin < mosquito < meal < bower < foot < shirt < step < child < stone < body < anchor < clothes < seed < exile < shippe < dinner < trench < element < lung < light < block < poet < sink < king < stair < breath < fool < phone < coward < banker < result < base < response < sip < bench < end < lock < victim < source < torrent < brick < sail < daughter < master < watch < gully < cross < scene < disciple < lady < food < direction < teacher < boy < middle < boat < messenger < parent < precipice < person < call < window < shore < wife < vessel < horse < temple < servant < piano < bed < patient < side < something < parcel < back < way < position < wall < place < lover < wind < state < corner < office < father < prison < worker < volunteer < street < abode < coach < flood < doorway < anything < someone < ground < front < brother < world < horseback < shop < current < city < energy < reservation < friend < camp < store < bank < factory < gentleman < rain < lad < deck < soul < home < beach < everything < floor < clock < car < house < door < ship < heaven < truck < air < system < barn < stream < mountain < restaurant < road < river < sea < bay < gate < hill < coast < farm < town < train < sun < room

Table A2: **Example of an Ordering Over all Words in a Set:** The trained weights of the linear probing model multiplied with the embedding of a word can help form a ‘global’ ordering over all the words (§4.1.2). This particular example is when the weight corresponding to the label 1 is used for the words in the Verb Physics size dev set with GloVe embeddings.