Topic Sensitive Attention on Generic Corpora Corrects Sense Bias in Pretrained Embeddings (Appendix)

Here we present additional results that did not fit into the main paper.

	# Words in target	Vocab size	# duplicate questions
Physics	542K	6,026	1981
Gaming	302K	6,748	3386
Android	235K	4,004	3190
Unix	262K	6,358	5312

Table 10: Statistics of the Stack Exchange data used in duplicate question detection. $\mathcal{D}_S$ has a vocabulary of 300,000 distinct words.

Method	Physics	Gamng	Andrd	Unix	Med
Tgt	121.9	185.0	142.7	159.5	158.9
SrcTune	$2.3_{\pm 0.7}$	$6.8_{\pm 0.3}$	$1.1_{\pm 0.3}$	$3.1_{\pm 0.0}$	$5.5_{\pm 0.8}$
RegFreq	$2.1_{\pm 0.8}$	$7.1_{\pm 0.7}$	$1.8_{\pm 0.4}$	$3.4_{\pm 0.5}$	$6.8_{\pm 0.9}$
RegSense	$5.0_{\pm 0.1}$	$13.8_{\pm 0.3}$	$6.7_{\pm 0.8}$	$9.7_{\pm 0.3}$	$14.6_{\pm 1.0}$
SrcSel	$5.8_{\pm 0.9}$	$11.7_{\pm 0.6}$	$5.9_{\pm 1.2}$	$6.4_{\pm 0.1}$	$8.6_{\pm 3.0}$
SrcSel +RegSense	$6.2_{\pm 1.3}$	$12.5_{\pm 0.3}$	$7.9_{\pm 1.8}$	$9.3_{\pm 0.2}$	$10.5_{\pm 0.9}$

Table 11: Average reduction in language model perplexity over Tgton five StackExchange Topics. ($\pm$ standard deviation) are shown in supplementary.

Ablation studies on SrcSel We compare variants in the design of SrcSel in Tables 12 and 13. In SrcSel:R we run the SrcSel without weighting the source snippets by the $Q(w, C)$ score in (6). We observe that the performance is worse than with the $Q(w, C)$ score. Next, we check if the score would suffice in down-weighting irrelevant snippets without help from our IR based selection. In SrcSel:c we include 5% random snippets from $\mathcal{D}_S$ in addition to those in SrcSel and weigh them all by their $Q(w, C)$ score. We find in Table 12 that the accuracy drops compared to SrcSel. Thus, both the $Q(w, C)$ weighting and the IR selection are important components of our source selection method.

	LM Perplexity				Question Dedup: AUC			
	Physics	Gaming	Android	Unix	Physics	Gaming	Android	Unix
Tgt	$121.9_{\pm 0.6}$	$185.0_{\pm 0.3}$	$142.7_{\pm 2.7}$	$159.5_{\pm 1.2}$	$86.7_{\pm 0.4}$	$82.6_{\pm 0.4}$	$86.8_{\pm 0.5}$	$85.3_{\pm 0.3}$
SrcSel:R	$114.8_{\pm 0.2}$	$172.7_{\pm 1.5}$	$131.6_{\pm 0.7}$	$151.8_{\pm 1.1}$	$89.2_{\pm 0.2}$	$85.6_{\pm 0.4}$	$87.5_{\pm 0.3}$	$86.8_{\pm 0.2}$
SrcSel:c					$88.7_{\pm 0.3}$	$84.8_{\pm 0.3}$	$87.0_{\pm 0.5}$	$85.8_{\pm 0.3}$
SrcSel	$116.1_{\pm 0.9}$	$173.3_{\pm 0.6}$	$136.7_{\pm 1.1}$	$153.1_{\pm 0.1}$	$90.4_{\pm 0.2}$	$85.4_{\pm 0.5}$	$87.4_{\pm 0.4}$	$87.5_{\pm 0.1}$

Table 12: Ablation studies on SrcSel over LM perplexity and AUC.

Critical hyper-parameters The number of neighbours K used for computing embedding based stability score as shown in (2) is set to 10 on all the tasks. We train each of the different embedding methods for a range of different epochs: $\{5, 20, 80, 160, 200, 250\}$. The λ parameter of RegSense and RegFreq is tuned over $\{0.1, 1, 10, 50\}$. Pre-trained embeddings $\mathcal{E}$ are obtained by training on CBOW model for 5 epochs on a cleaned version of 20160901 dump of Wikipedia. All the embedding sizes irrespective of the training method are set to 300.

	Micro Accuracy	Macro Accuracy
Tgt	$26.3_{\pm 0.5}$	$14.7_{\pm 1.2}$
SrcSel:R	$27.3_{\pm 0.3}$	$16.1_{\pm 1.6}$
SrcSel	$28.3_{\pm 0.4}$	$17.3_{\pm 0.7}$

Table 13: Ablation studies on SrcSel for classification on the Medical dataset.

	Perplexity				AUC			
	Physics	Gaming	Android	Unix	Physics	Gaming	Android	Unix
Tgt	121.9	185.0	142.7	159.5	86.7	82.6	86.8	85.3
RegFreq	$2.1_{\pm 0.8}$	$7.0_{\pm 0.7}$	$1.8_{\pm 0.4}$	$3.4_{\pm 1.0}$	$-0.4_{\pm 0.4}$	$2.3_{\pm 0.3}$	$-0.6_{\pm 0.4}$	$-0.3_{\pm 0.3}$
RegFreq-rinit	$-1.6_{\pm 0.9}$	$1.2_{\pm 0.6}$	$1.6_{\pm 0.2}$	$2.6_{\pm 0.8}$	$-1.2_{\pm 0.2}$	$0._{\pm 0.3}$	$-0.2_{\pm 0.5}$	$-0.3_{\pm 0.3}$
RegSense	$5.0_{\pm 0.1}$	$13.8_{\pm 0.3}$	$6.7_{\pm 0.7}$	$9.7_{\pm 0.8}$	$-0.3_{\pm 0.3}$	$2.1_{\pm 0.4}$	$-0.6_{\pm 0.3}$	$-0.3_{\pm 0.5}$
RegSense-rinit	$3.6_{\pm 1.0}$	$11.1_{\pm 0.5}$	$7.0_{\pm 1.2}$	$8.9_{\pm 1.4}$	$0.7_{\pm 0.2}$	$1.2_{\pm 0.2}$	$-0.3_{\pm 0.6}$	$-0.2_{\pm 0.5}$
SrcSel	$5.8_{\pm 0.9}$	$11.7_{\pm 1.8}$	$6.0_{\pm 1.2}$	$6.3_{\pm 1.0}$	$3.7_{\pm 0.2}$	$2.8_{\pm 0.5}$	$0.6_{\pm 0.4}$	$2.2_{\pm 0.1}$
SrcSel:R-rinit	$5.8_{\pm 1.0}$	$12.5_{\pm 0.4}$	$10.4_{\pm 1.4}$	$7.9_{\pm 1.0}$	$2.5_{\pm 0.1}$	$2._{\pm 0.5}$	$0.4_{\pm 0.4}$	$1.5_{\pm 0.2}$

Table 14: Source initialization vs random initialization (with suffix *-rinit*) on RegFreq , RegSense , SrcSel . Shown in the table is the average gain over Tgt for each method.

	Medical			20 Newsgroup Micro				
Method	Micro	Macro	Rare	Science	Computer	Politics	Religion	Recreation
Tgt	26.3	14.7	3.0	92.2	79.9	94.8	87.3	90.3
Src	$-1.0_{\pm 0.9}$	$0.0_{\pm 0.5}$	$0.0_{\pm 0.1}$	$-0.1_{\pm 0.8}$	$-9.1_{\pm 0.7}$	$-3.3_{\pm 0.9}$	$-1.0_{\pm 2.3}$	$-6.0_{\pm 0.5}$
SrcTune	$1.7_{\pm 1.0}$	$1.8_{\pm 1.7}$	$1.5_{\pm 2.0}$	$0.0_{\pm 1.1}$	$0.0_{\pm 1.3}$	$-0.1_{\pm 1.5}$	$0.1_{\pm 2.5}$	$0.2_{\pm 1.1}$
RegSense	$1.4_{\pm 0.5}$	$2.5_{\pm 1.2}$	$4.0_{\pm 1.8}$	$0.9_{\pm 0.9}$	$-0.2_{\pm 1.3}$	$0.2_{\pm 1.3}$	$1.2_{\pm 2.0}$	$0.1_{\pm 0.8}$
SrcSel	$2.0_{\pm 0.9}$	$2.6_{\pm 1.5}$	$1.1_{\pm 1.4}$	$1.2_{\pm 0.8}$	$0.1_{\pm 1.4}$	$0.5_{\pm 1.2}$	$0.5_{\pm 2.6}$	$0.3_{\pm 1.0}$

Table 15: Average accuracy gains over Tgt and ±std-dev on six classification domains across Medical and the 20 NG datasets. Medical classes have more skew, so also showing macro and rare class accuracy gains.

Method	Physics	Gaming	Android	Unix	Med (Micro)	Med (Rare)
Tgt	$89.7_{\pm 0.3}$	$88.4_{\pm 0.2}$	$89.4_{\pm 0.3}$	$89.2_{\pm 0.3}$	$31.4_{\pm 0.9}$	$9.4_{\pm 3.0}$
SrcTune	$89.5_{\pm 0.2}$	$89.0_{\pm 0.2}$	$89.0_{\pm 0.2}$	$89.0_{\pm 0.2}$	$31.3_{\pm 0.4}$	$7.3_{\pm 2.6}$
SrcSel	$91.6_{\pm 0.3}$	$88.9_{\pm 0.8}$	$89.4_{\pm 0.2}$	$89.0_{\pm 0.2}$	$31.3_{\pm 0.9}$	$10.5_{\pm 1.8}$

Table 16: Performance with a larger target corpus size of 10MB on the four deduplication tasks (AUC score) and one classification task (Accuracy) shown in the last two columns.

	Physics	Gaming	Android	Unix	Med
Tgt	86.7	82.6	86.8	85.4	26.3
Elmo	$-1.0_{\pm 0.4}$	$4.5_{\pm 0.3}$	$-1.5_{\pm 0.8}$	$-2.3_{\pm 0.3}$	$3.2_{\pm 1.3}$
+Tgt	$-0.8_{\pm 0.4}$	$3.8_{\pm 0.4}$	$0.5_{\pm 0.1}$	$-0.0_{\pm 0.1}$	$4.1_{\pm 1.5}$
+SrcTune	$-0.5_{\pm 0.3}$	$3.0_{\pm 0.2}$	$0.3_{\pm 0.5}$	$0.2_{\pm 0.2}$	$3.5_{\pm 0.6}$
+SrcSel	$2.6_{\pm 0.5}$	$4.1_{\pm 0.1}$	$1.1_{\pm 0.4}$	$1.5_{\pm 0.2}$	$4.6_{\pm 0.9}$

Table 17: AUC scores comparing contextual embeddings on the question dedup tasks and Medical Abstract Classification task. Shown in the first row is the performance of Tgt and the rows below show mean gains over Tgt (± std-dev).