A Rules and Examples

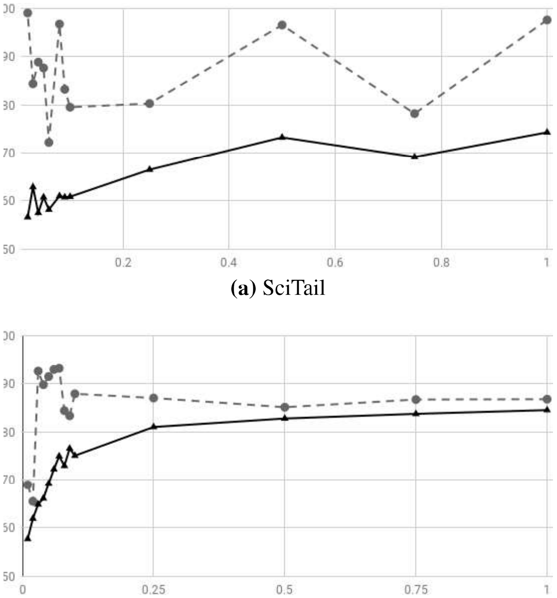
Table 8: Number of rules in $\mathbb{G}^{\mathrm{KB}}$

KB	PPDB	SICK	WordNet
#Rules	6,977,679	12,511	∼116,000
Examples	because of ⇒ due to, wish ⇒ would like	woods ⇒ wooden area, kid ⇏ woman	car ⇒ cabin car, hate ⇏ love

Table 8 shows the number of rules and additional examples for $\mathbb{G}^{\mathrm{KB}}$.

B Training data sizes

Figure 3 shows training (dotted) accuracies on sub-sampled training datasets and testing (solid) accuracies on original test dataset X $_{test}$ of $\mathbb{D}$ over different sub-sampling percentages of the training set. Since SciTail (27K) is much smaller than SNLI (570K), SciTail fluctuates a lot at smaller sub-samples while SNLI converges with just 50% of the examples.



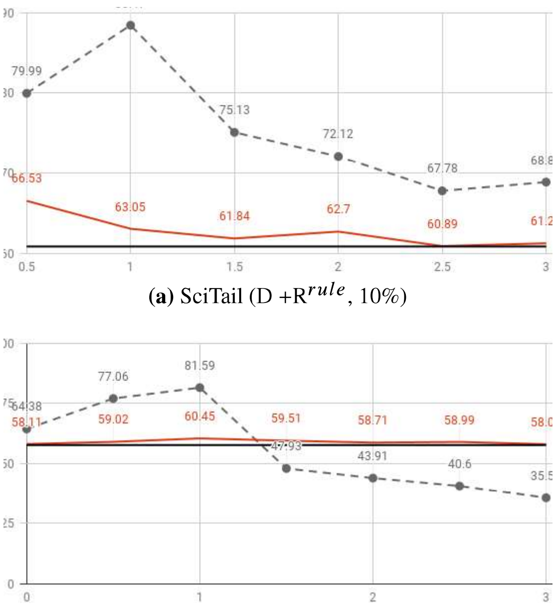
(a) SciTail

(b) SNLI

Figure 3: $\mathbb{D}$ for SciTail and SNLI.

C Effectiveness of Z/X Ratio, α

Figure 4 shows train/test accuracies with different balancing ratio between z and x. The dotted line is training accuracies, the solid black horizontal line is testing accuracy of $\mathbb{D}$. The solid red shows test accuracies with different balancing ratio, α (x-axis) from 0.5, 1.0, ... 3.0 from $|z| = \alpha * |x|$ where $|x|$ is fixed as batch size. The generated examples z are useful up to a point, but the performance quickly degrades for $\alpha > 1.0$ as they overwhelm the original dataset x.



(a) SciTail (D +R^{rule}, 10%)

(b) SNLI (D +R^{rule}, 1%)

Figure 4: Effect of balancing ratio between z and x.

D Retrofitting Experiment

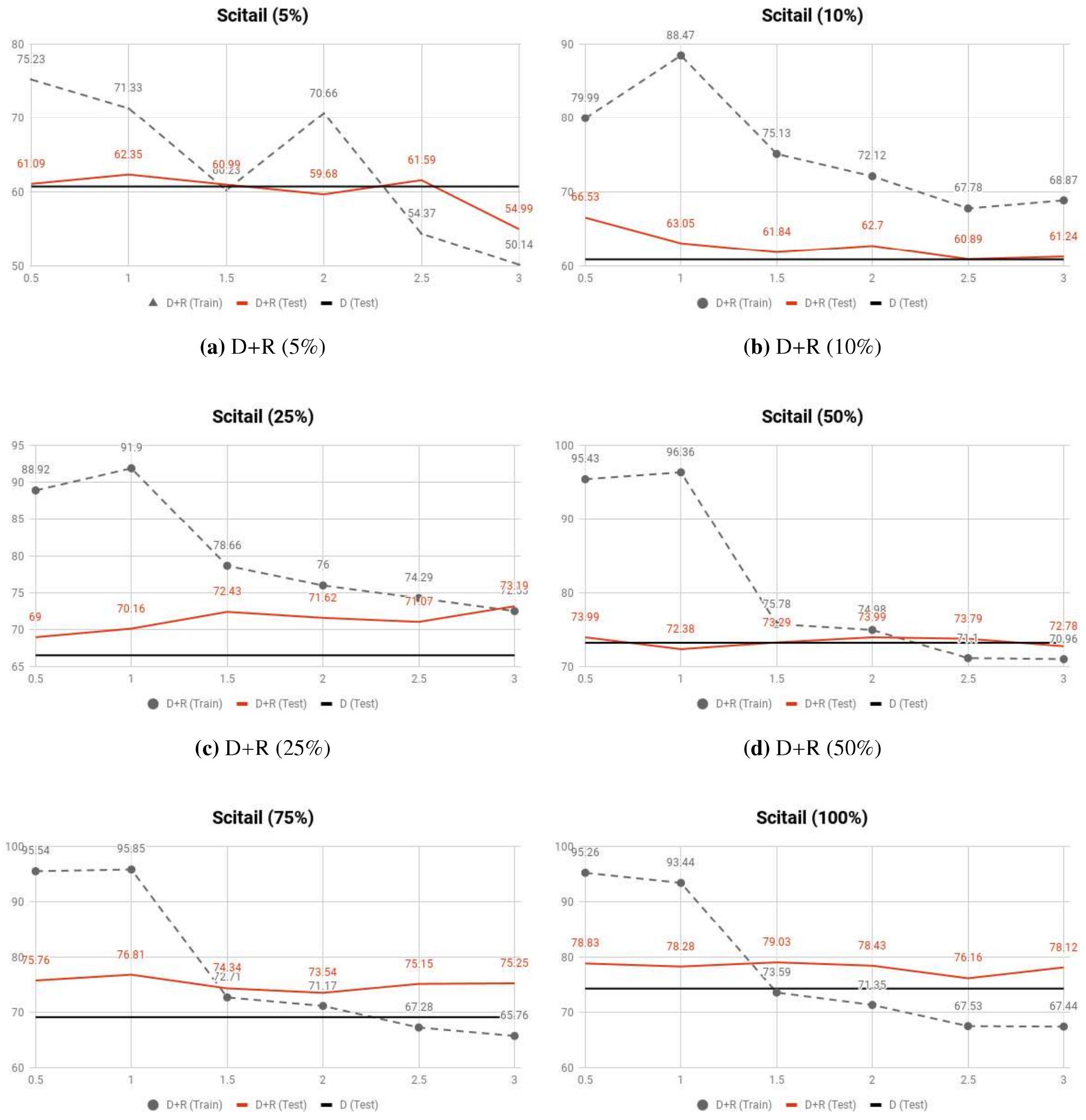
Table 9 shows the grid search results of retrofitting vectors (Faruqui et al., 2015) with different lexical resources. To obtain the strongest baseline, we choose the best performing vectors for each sub-sample ratio and each dataset. Usually, PPDB and WordNet are two most useful resources for both SNLI and SciTail.

E In-Depth Analysis: D+R

Table 5 and Table 6 show more in-depth analysis with different sub-sampling ratio on SNLI and SciTail. The dotted line is training accuracy, and the solid red ($\mathbb{D}$ +$\mathbb{G}^{\mathrm{rule}}$) and sold black ($\mathbb{D}$) shows testing accuracies.

Table 9: Results of the word vectors retrofitted on different lexicons on each dataset. We pick the best vectors for each task and sub-sampling ratio.

ratio	Lexicon	SNLI	SciTail
1%	framenet	56.15	60.89
1%	ppdb	**57.04**	**62.5**
1%	wordnet	55.58	62.2
1%	all	56.81	61.14
10%	framenet	72.75	**67.99**
10%	ppdb	72.88	54.74
10%	wordnet	73.27	67.29
10%	all	**73.45**	66.43
50%	framenet	80.95	66.08
50%	ppdb	81.14	67.24
50%	wordnet	80.62	**69.05**
50%	all	**81.18**	68.4
100%	framenet	83.66	70.06
100%	ppdb	**84.14**	70.16
100%	wordnet	83.91	**72.63**
100%	all	83.68	71.12



(a) D+R (5%)

(b) D+R (10%)

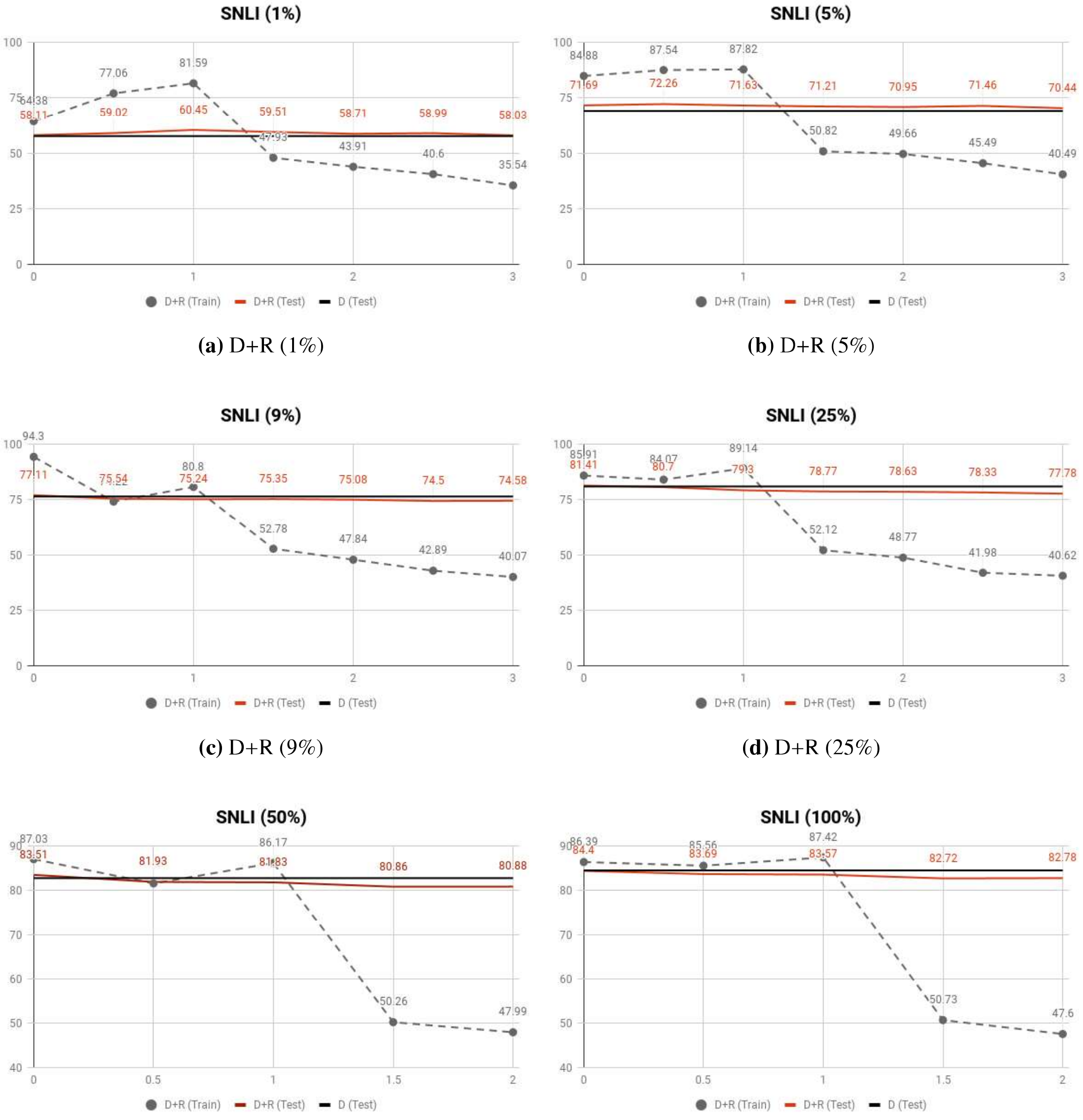
(c) D+R (25%)

(d) D+R (50%)

(e) D+R (75%)

(f) D+R (100%)

Figure 5: $\mathbb{D}$ +$\mathbb{G}^{\text{rule}}$ with different ratio for SciTail.



(**a**) D+R (1%)

(**b**) D+R (5%)

(**c**) D+R (9%)

(**d**) D+R (25%)

(**e**) D+R (50%)

(**f**) D+R (100%)

Figure 6: $\mathbb{D}$ +$\mathbb{G}^{\text{rule}}$ with different ratio for SNLI.