

Appendix

Lic.	NPI	Sentence
1	1	Has the guy worked with <i>any</i> teenagers?
1	0	Has the guy worked with <i>the</i> teenagers?
0	1	*The guy has worked with <i>any</i> teenagers.
0	0	The guy has worked with <i>the</i> teenagers.

Table 3: Reduced paradigm for Simple questions. ‘Lic.’ is abbreviated from ‘Licensor’. The licensor and licensor replacement are shown in bold (*has* in both cases). The NPI (*any*) and NPI replacement (*the*) are shown in italics. There is no scope manipulation because it is not possible to place an NPI or NPI replacement outside of the scope of an interrogative or declarative phrase. The 2 minimal pairs are shown by arrows, pointing from unacceptable to acceptable sentence.

Environment	Label	% accept	Diff
Adverb	*	8.33	61.67
	✓	70.00	
Conditionals	*	37.50	50.00
	✓	87.50	
Determiner negation	*	11.11	78.89
	✓	90.00	
Embedded questions	*	8.33	89.17
	✓	97.50	
Only	*	5.56	84.44
	✓	90.00	
Sentential negation	*	27.78	52.22
	✓	80.00	
Simple questions	*	33.33	62.97
	✓	96.30	
Superlatives	*	8.33	66.67
	✓	75.00	
Quantifiers	*	4.17	50.83
	✓	55.00	

Table 4: Results from MTurk validation. ‘Environment’ is the name of the licensing environment and ‘label’ is whether the sentence was intended as acceptable (✓) or unacceptable (*). The results of the validation ratings is in ‘% accept’ and represents the majority vote for each sentence as acceptable/unacceptable and then averaged to give the percentage of times a sentence in a given condition was rated as acceptable by the MTurk raters. ‘Diff’ is calculated from the % of acceptable sentences rated acceptable minus the % of unacceptable sentences rated acceptable (100 is a perfect score, 0 means there is no difference).

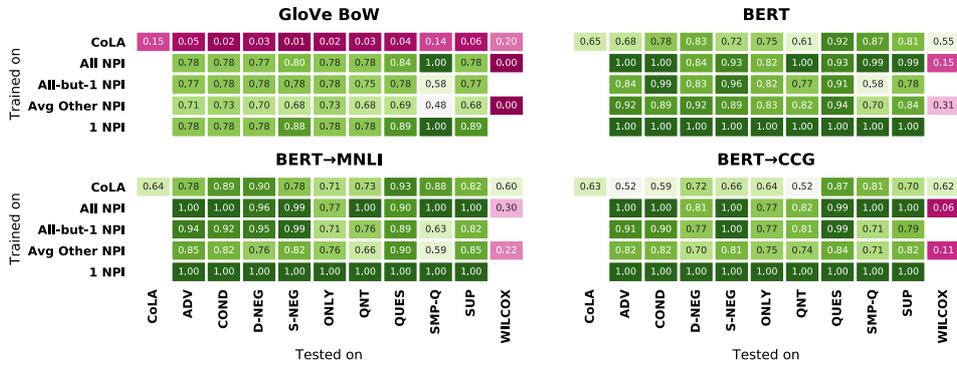


Figure 5: Results from the acceptability judgment experiment in MCC. The columns indicate evaluation tests, and the rows fine-tuning settings.

NPI-Present: 1-0



Figure 6: Results from minimal pair test for the NPI-presence contrast. The smaller diagrams of each sector show performance of BoW and BERT variants under two different minimal pair evaluation methods. The rows represent training-evaluation configuration, while the columns represent different licensing environments.

Licensor: 0-1, Scope: 1, NPI-Present: 1

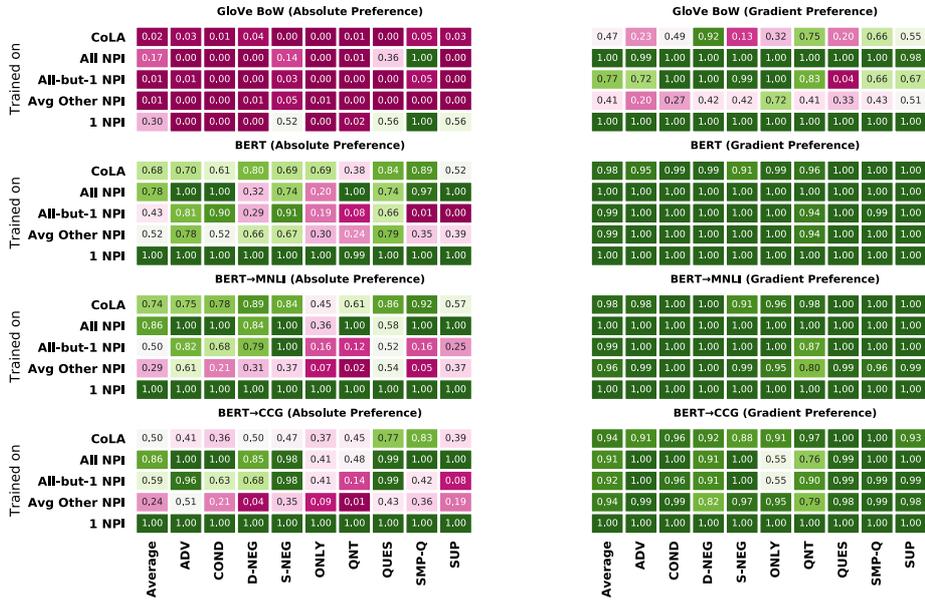


Figure 7: Results from minimal pair test for the licensor-presence contrast. The smaller diagrams of each sector show performance of BoW and BERT variants under two different minimal pair evaluation methods. The rows represent training-evaluation configuration, while the columns represent different licensing environments.

Licensor: 1, Scope: 0-1, NPI-Present: 1



Figure 8: Results from minimal pair test for the scope contrast. The smaller diagrams of each sector show performance of BoW and BERT variants under two different minimal pair evaluation methods. The rows represent training-evaluation configuration, while the columns represent different licensing environments.

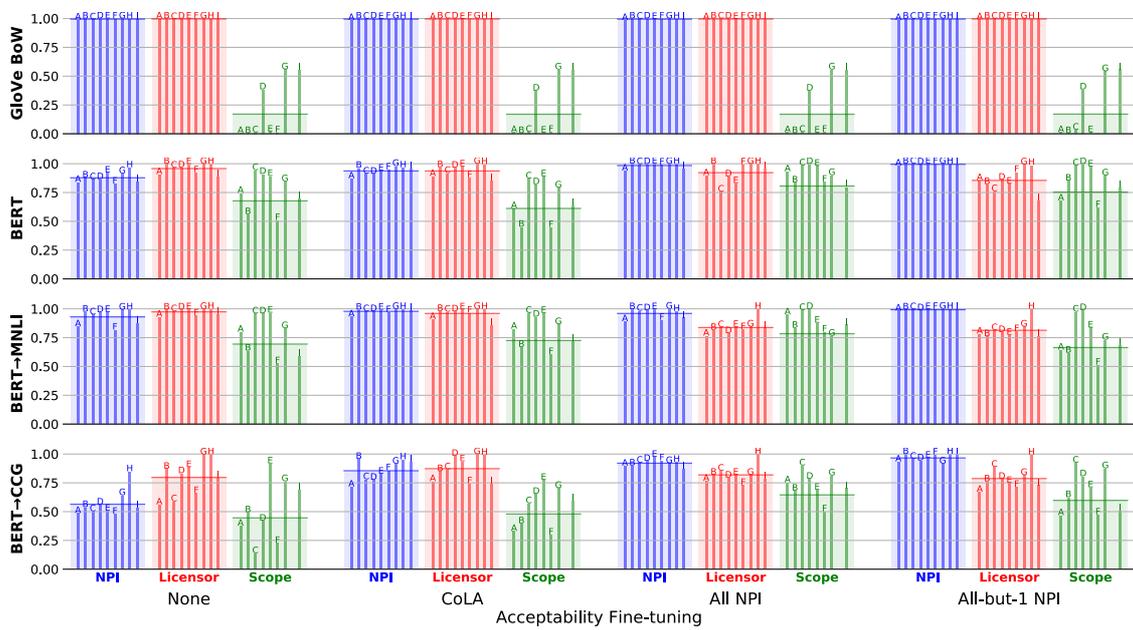


Figure 9: Results of probing classification on NPI presence, licenser presence, and scope detection, shown in MCC. Letters on top of bars refer to NPI environments: A=ADV, B=COND, C=D-NEG, D=S-NEG, E=ONLY, F=QNT, G=QUES, H=SMP-Q, I=SUP.