

Model	Noise	Annotated Evidence									Retrieved Evidence		
		Strict (Val.)			Strict (Test)			Fuzzy (Test)			Fuzzy (Test, Vote)		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
Softmax	False	58.21	72.90	64.73	57.42	72.32	64.02	61.53	77.49	68.59	67.53	74.18	70.85
Softmax ($k=1$)	False	58.28	71.80	64.34	58.22	71.83	64.31	62.48	77.08	69.02	67.06	73.47	70.26
CRF	False	63.19	79.21	70.30	61.90	77.33	68.76	66.00	82.45	73.32	68.97	74.64	71.80
Softmax	True	59.74	69.11	64.08	59.38	68.77	63.73	63.58	73.63	68.24	69.75	74.72	72.23
Softmax ($k=1$)	True	59.84	67.51	63.44	59.76	67.61	63.44	64.02	72.44	67.97	69.11	73.93	71.51
CRF	True	64.42	75.84	69.67	63.72	76.09	69.36	67.53	80.63	73.50	72.66	76.83	74.74