Appendix

A Crowdsourcing Details

Crowd workers residing in five English-speaking countries (United States, United Kingdom, New Zealand, Australia and Canada) were hired. Each crowd worker had a Level 2 or higher rating on Figure Eight, which corresponds to a "group of more experienced, higher accuracy contributors". Each contributor had to pass a test questionnaire to be eligible to take part in the experiment. Test questions were also hidden throughout the task and untrusted contributions were removed from the final dataset. For greater quality control, an upper limit of 75 judgments per contributor was enforced.

Crowd workers were paid a total of $1 for 50 judgments. An internal unpaid workforce (including the first and second author of the paper) of 7 contributors was used to speed up data collection.

# Judgments			Average	Sentence
Positive	Negative	Neutral		
1	1	7	0.50	the fight scenes are fun , but it grows tedious
3	2	4	0.56	it 's not exactly a gourmet meal but the fare is fair , even coming from the drive thru
2	3	4	0.44	propelled not by characters but by caricatures
4	2	3	0.61	not everything works , but the average is higher than in mary and most other recent comedies

Table A1: Examples of neutral sentences for a threshold of 0.66

# Judgments			Average	Original	Sentence
Positive	Negative	Neutral			
1	5	3	0.28	Positive	de niro and mcdormand give solid performances , but their screen time is sabotaged by the story 's inability to create interest
6	0	3	0.83	Negative	son of the bride may be a good half hour too long but comes replete with a flattering sense of mystery and quietness
0	5	4	0.22	Positive	wasabi is slight fare indeed , with the entire project having the feel of something tossed off quickly (like one of hubert 's punches) , but it should go down smoothly enough with popcorn

Table A2: Examples of flipped sentiment sentences, for a threshold of 0.66

Model 1		vs	Model 2		Significant
distill	no-project		distill	project	Yes
no-distill	no-project		no-distill	project	Yes
ELMo	no-project		ELMo	project	No
no-distill	no-project		distill	no-project	No
no-distill	project		distill	project	No
no-distill	no-project		ELMo	no-project	Yes
distill	no-project		ELMo	no-project	Yes
no-distill	project		ELMo	project	Yes
distill	project		ELMo	project	Yes

Table A3: Statistical significance using a two-sided Kolmogorov-Smirnov statistic (Massey Jr, 1951) with $\alpha = 0.001$.

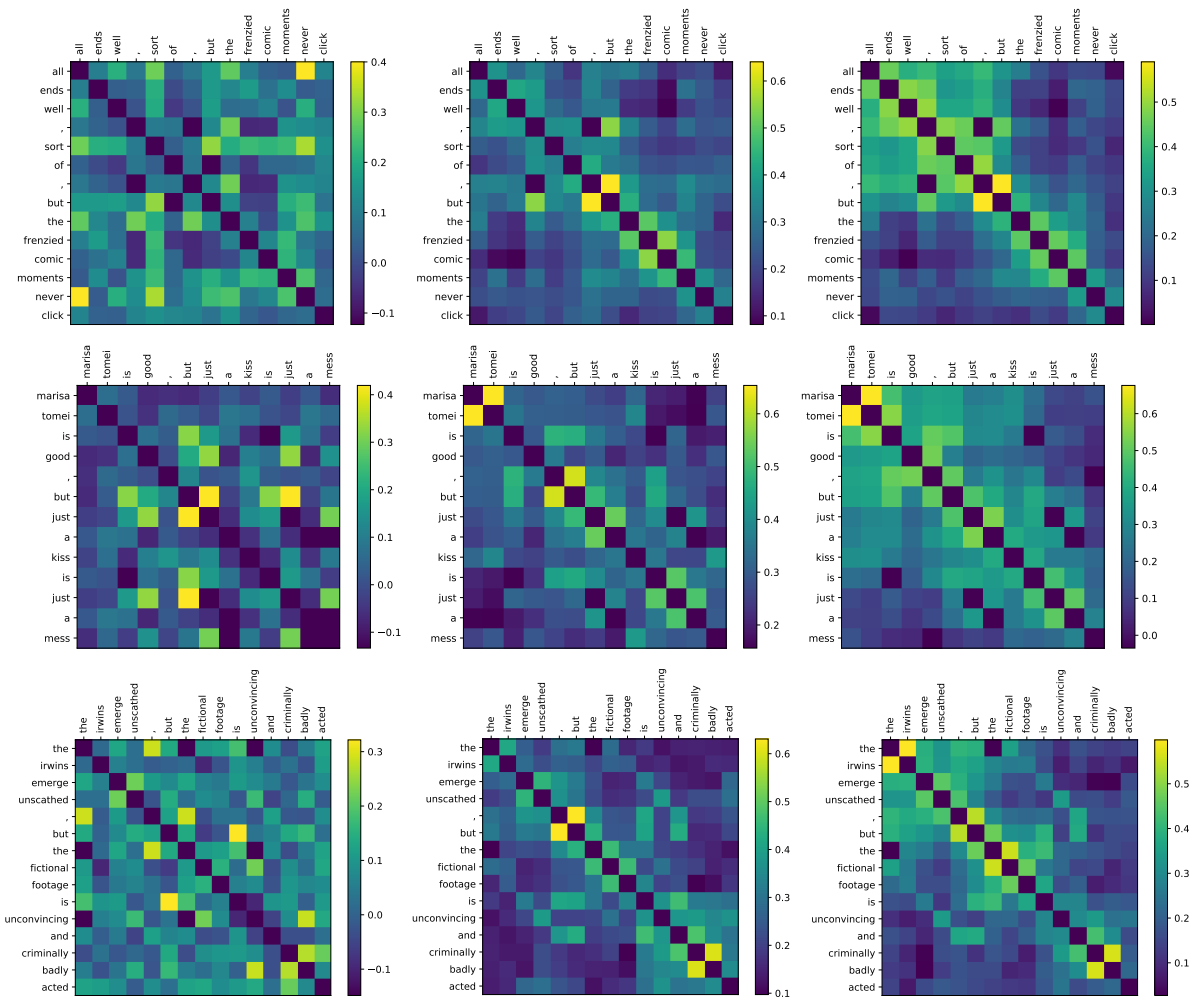

Figure A1: Heat map showing the cosine similarity between pairs of word vectors within a single sentence. The leftmost column has `word2vec` (Mikolov et al., 2013) embeddings, fine-tuned on the downstream task (SST2). The middle column contains the original ELMo embeddings (Peters et al., 2018a) without any fine-tuning. The representations from the three layers (token layer and two LSTM layers) have been averaged. The rightmost column contains ELMo embeddings fine-tuned on the downstream task. For better visualization, the cosine similarity between identical words has been set equal to the minimum value in the map.