A Appendix: Additional Results

System	user -entity	mention -entity	NEEL-test			TACL			Avg. F1
			P	R	F1	P	R	F1	
Our approach									
NTEL-nonstruct			83.0	71.8	77.0	80.9	69.0	74.5	75.8
NTEL			84.4	73.9	78.8	82.0	71.3	76.3	77.6
NTEL	✓		83.8	76.7	80.1	81.8	**73.3**	77.3	78.7
NTEL		✓	84.1	78.3	81.1	83.0	71.7	76.9	79.0
NTEL	✓	✓	**84.8**	**79.3**	**82.0**	**83.5**	72.7	**77.7**	**79.9**
Best published results									
S-MART			83.2	79.2	81.1	76.8	73.0	74.9	78.0

Table 6: Evaluation results on the NEEL-test and TACL datasets for different systems. Twitter messages that contain no ground truth entities are excluded for both training and testing. The best results are in **bold**.

In the first version of (Yang and Chang, 2015), the Twitter messages that contain no ground truth entities are excluded in the experiments. For completeness, we now present the evaluation results of NTEL in this setting, which are shown in Table 6. The RETWEET+ network is adopted to train author embeddings. The best hyper-parameters are the same as those described in § 5, except for the L2 regularization penalty for the composition parameters, which is set as 0.01 here.

The results are generally better than those presented in Table 4. As shown, NTEL benefits from the distributed representations of authors, mentions, and entities, which improve the average F1 score by 2.3 points. NTEL also gives the best results on the datasets, outperforming S-MART by about 2% F1 on average.