Improved Entity Tracking Supplementary Material

Model Hyperparameters In addition to those described in the main paper, we use the following hyperparameters in our experiments, which we obtained via random search:

	dim	clip	dropout	β_1	η	ent types	ent tokens
Lamb, AttSum-Feat	300	10	0.4	0.9	0.0005	-	-
Lamb, AttSum-Feat + L^1	100	10	0.2	0.7	0.001	5	100
Lamb, AttSum-Feat + L^2	128	1	0.4	0.7	0.001	2	2
CBT-NE, AttSum-Feat	300	10	0.1	0.7	0.0005	-	-
CBT-NE, AttSum-Feat + L^1	256	10	0.2	0.9	0.001	3	3
CBT-NE, AttSum-Feat + L^2	256	10	0.5	0.9	0.001	2	2

We elaborate on these below:

1. dim: The dimensionality of word embeddings and RNN states
2. clip: Gradients were rescaled to not exceed this value in norm
3. dropout: Dropout rate
4. β_1: ADAM hyperparameter
5. η ADAM learning rate
6. ent types: distinct named entity word types allowed in multi-task loss
7. ent tokens: named entity word tokens used in multi-task loss

We used a batch-size of 64 in all experiments, initialized all parameters to lie uniformly in $[-0.1, 0.1]$, and set $\gamma = 0.5$ for all multi-task experiments. We used (at most) the last 1024 tokens in each LAMBADA example in defining x, and at most the last 1500 tokens in each CBT-NE example in defining x.

Training Details We sort the examples by length in descending order and the mini-batches are taken as continuous chunks from this set (but at random index in each epoch).

Speaker Id To heuristically determine the speaker, we use the following pseudo-code rules:

```
if a quote doesn't end with a '.':
    if there is a PERSON w/ in the 10 tokens following the end of the quote:
        the speaker is the closest PERSON following the end of the quote
    else:
        the speaker is the closest PERSON that precedes the beginning of the quote
else:
    the speaker is the closest PERSON that precedes the beginning of the quote
```

Statistical Significance Test We use McNemar's test on 3 different types of comparisons and list the p-values obtained below (values greater than 0.05 are highlighted in red):

1. AttSum* vs AttSum, in table 1
2. AttSum-Feat + $\mathcal{L}^i$ vs AttSum + $\mathcal{L}^i$, in table 2
3. AttSum-Feat + $\mathcal{L}^i$ vs AttSum-Feat, in table 3

LAMBADA	Val	Test
AttSum + $\mathcal{L}^1$	1.43e-05	1.44e-02
AttSum + $\mathcal{L}^2$	1.27e-04	1.66e-03
AttSum-Feat	8.93e-11	3.59e-11
AttSum-Feat + $\mathcal{L}^1$	1.49e-13	1.58e-11
AttSum-Feat + $\mathcal{L}^2$	1.17e-12	2.83e-07
CBT-NE		
AttSum + $\mathcal{L}^1$	3.91e-02	1.02e-02
AttSum + $\mathcal{L}^2$	3.72e-03	1.09e-03
AttSum-Feat	9.78e-05	4.66e-03
AttSum-Feat + $\mathcal{L}^1$	3.52e-06	1.54e-07
AttSum-Feat + $\mathcal{L}^2$	3.37e-08	5.53e-03

Table 1: p-values for performance comparison against AttSum

LAMBADA	Val	Test
AttSum-Feat + $\mathcal{L}^1$	7.91e-04	5.49e-06
AttSum-Feat + $\mathcal{L}^2$	2.43e-04	3.11e-02
CBT-NE		
AttSum-Feat + $\mathcal{L}^1$	1.51e-02	7.07e-03
AttSum-Feat + $\mathcal{L}^2$	3.37e-03	8.49e-01

Table 2: p-values for performance comparison of AttSum-Feat + $\mathcal{L}^i$ against AttSum + $\mathcal{L}^i$

LAMBADA	Val	Test
AttSum-Feat + $\mathcal{L}^1$	2.67e-01	7.51e-01
AttSum-Feat + $\mathcal{L}^2$	3.57e-01	2.71e-01
CBT-NE		
AttSum-Feat + $\mathcal{L}^1$	4.93e-01	7.42e-03
AttSum-Feat + $\mathcal{L}^2$	5.90e-02	1.0

Table 3: p-values for performance comparison of AttSum-Feat + $\mathcal{L}^i$ against AttSum-Feat