A Model parameters

Table 3 shows the hyperparameters used to train the models. We have trained each model up to 100 iterations and kept the one with the highest LAS score on the development set. The models were optimized with Stochastic Gradient Descent (SGD) with a batch size of 8. During multitask learning, dependency parsing as the main task was weighted 1.0 while the gaze data treated as auxiliary task had a weight of 0.1.

Initial learning rate	0.02
Time-based learning rate decay	0.05
Momentum	0.9
Dropout	0.5
DIMENSIONS	
Word embedding	100
Char embedding	30
Self-defined features	20
Word hidden vector	800
Character hidden vector	50

Table 3: Model hyperparameters.