Model	English F_1	Chinese F_1
Full Model	65.52	64.41
– MENTION	–1.27	–0.74
– GENRE	–0.25	–2.91
– DISTANCE	–2.42	–2.41
– SPEAKER	–1.26	–0.93
– MATCHING	–2.07	–3.44

Table 1: CoNLL F_1 scores of the mention-ranking model on the dev sets without mention, document genre, distance, speaker, and string matching hand-engineered features.

6 Experiments and Results

Experimental Setup. We run experiments on the English and Chinese portions of the CoNLL 2012 Shared Task data (Pradhan et al., 2012). The models are evaluated using three of the most popular coreference metrics: MUC, B^3, and Entity-based CEAF ($CEAF_{\phi_4}$). We generally report the average F_1 score (CoNLL F_1) of the three, which is common practice in coreference evaluation. We used the most recent version of the CoNLL scorer (version 8.01), which implements the original definitions of the metrics.

Mention Detection. Our experiments were run using system-produced predicted mentions. We used the rule-based mention detection algorithm from Raghunathan et al. (2010), which first extracts pronouns and maximal NP projections as candidate mentions and then filters this set with rules that remove spurious mentions such as numeric entities and pleonastic *it* pronouns.

6.1 Mention-Ranking Model Experiments

Feature Ablations. We performed a feature ablation study to determine the importance of the hand-engineered features included in our model. The results are shown in Table 1. We find the small number of non-embedding features substantially improves model performance, especially the distance and string matching features. This is unsurprising, as the additional features are not easily captured by word embeddings and historically such features have been very important in coreference resolvers (Bengtson and Roth, 2008).

The Importance of Pretraining. We evaluate the benefit of the two-step pretraining for the mention-ranking model and report results in Table 2. Consistent with Wiseman et al. (2015), we find pretraining to greatly improve the model's accuracy. We note in particular that the model benefits from using both pretraining steps from Section 4, which more smoothly transitions the model from a mention-pair classification objective that is easy to optimize to a max-margin objective better suited for a ranking task.

All-Pairs	Top-Pairs	English F_1	Chinese F_1
Yes	Yes	65.52	64.41
Yes	No	–0.36	–0.24
No	Yes	–0.54	–0.33
No	No	–3.58	–5.43

Table 2: CoNLL F_1 scores of the mention-ranking model on the dev sets with different pretraining methods.

Model	English F_1	Chinese F_1
Full Model	66.01	64.86
– PRETRAINING	–5.01	–6.85
– EASY-FIRST	–0.15	–0.12
– L2S	–0.32	–0.25

Table 3: CoNLL F_1 scores of the cluster-ranking model on the dev sets with various ablations. – PRETRAINING: initializing model parameters randomly instead of from the mention-ranking model, – EASY-FIRST: iterating through mentions in order of occurrence instead of according to their highest scoring candidate coreference link, – L2S: training on a fixed trajectory of correct actions instead of using learning to search.

6.2 Cluster-Ranking Model Experiments

We evaluate the importance of three key details of the cluster ranker: initializing it with the mention-ranking model's weights, using an easy-first ordering of mentions, and using learning to search. The results are shown in Table 3.

Pretrained Weights. We compare initializing the cluster-ranking model randomly with initializing it with the weights learned by the mention-ranking model. Using pretrained weights greatly improves performance. We believe the cluster-ranking model has difficulty learning effective weights from scratch due to noise in the signal coming from cluster-level decisions (an overall bad cluster merge may still involve a few cor-