Appendix 1: Experiments on ACE 2005 where Gold Entity Types Are Unknown

Experimental Settings: For comparison with prior work (Plank and Moschitti, 2013), we (1) generate relation instances from all pairs of entities within each sentence with three or fewer intervening entity mentions—labeling those pairs with no relation as negative instances, (2) use gold entity spans (but not types) at train and test time, and (3) evaluate on the 7 coarse relation types, ignoring the subtypes. In the training set, 35,990 total relations are annotated of which only 3,658 are non-nil relations. We did not match the number of tokens they reported in the `cts` and `wl` domains. Therefore, in this section we only report the results on the test set of `bc` domain. We will leave experiments on additional domains in future work.

We run the same models as in §7 on this task. Here the FCM does not use entity type features. Plank and Moschitti (2013) also use Brown clusters and word vectors learned by latent-semantic analysis (LSA). In order to make a fair comparison with their method, we also report the FCM result using Brown clusters (prefixes of length 5) of entity heads as entity types. Furthermore, we report non-comparable settings using WordNet super-sense tags of entity heads as types. The WordNet features were also used in their paper but not as substitution of entity types. We use the same toolkit to get the WordNet tags as in §6. The Brown clusters are from (Koo et al., 2008)[11].

Results: Table 7 shows the results under the low-resource setting. When no entity types are available, the performance of our FCM only model greatly decreases to 48.15%, which is consistent with our observation in the ablation tests. The baseline model also relies heavily on the entity types. After we remove all the hand-engineering features that contain entity type information, the performance of our baseline model drop to 40.62%, even lower than the reduced FCM only model.

The combination of baseline model and head embeddings (Baseline + HeadOnly) greatly improve the results. This is consistent with the observation in Nguyen and Grishman (2014) that when the gold entity types are unknown, information of the entity heads provided by their embeddings will play a more important role. Combination of the baseline and FCM (Baseline + FCM) also achieves improvement but not significantly better than Baseline + HeadOnly. A possible explanation is that FCM becomes less efficient on using context word embeddings when the entity type information is unavailable. In this situation the head embeddings provided by FCM become the dominating contribution to the baseline model, making the model have similar behavior as the Baseline + HeadOnly method.

Finally, we find Brown clusters can help FCM when entity types are unknown. Although the performance is still not significantly better than Baseline + HeadOnly, it outperforms all the results in Plank and Moschitti (2013) as a *single model*, and with the *same source of features*. WordNet supersense tags further improves FCM, and achieves the best reported results on this low-resource setting. These results are encouraging since it shows FCM may be more useful under the end-to-end setting where predictions of both entity mentions and relation mentions are required in place of predicting relation based on gold tags (Li and Ji, 2014).

Recently Nguyen et al. (2015) proposed a novel way of applying embeddings to tree-kernels. From the results, our best single model achieves comparable result with their best single system, while their combination method is slightly better than ours. This suggests that we may benefit more from combining the usages of multiple word representations; and we will investigate it in future work.

	bc		
Model	**P**	**R**	**F1**
PM'13 (Brown)	54.4	43.4	48.3
PM'13 (LSA)	53.9	45.2	49.2
PM'13 (Combination)	55.3	43.1	48.5
(1) FCM only	53.7	43.7	48.2
(3) Baseline	59.4	30.9	40.6
(4) + HeadOnly	64.9	41.3	50.5
(5) + FCM	**65.5**	41.5	50.8
(1) FCM only w/ Brown	64.6	40.2	49.6
(1) FCM only w/WordNet	64.0	43.2	51.6
Linear+Emb	46.5	**49.3**	47.8
Tree-kernel+Emb (Single)	57.6	46.6	51.5
Tree-kernel+Emb (Combination)	58.5	47.3	**52.3**

Table 7: Comparison of models on ACE 2005 out-of-domain test sets for the low-resource setting, where the gold entity spans are known but entity types are unknown. PM'13 is the results reported in Plank and Moschitti (2013). "Linear+Emb" is the implementation of our method (4) in (Nguyen et al., 2015). The "Tree-kernel+Emb" methods are the enrichments of tree-kernels with embeddings proposed by Nguyen et al. (2015).

[11] `http://people.csail.mit.edu/maestro/papers/bllip-clusters.gz`