System	Four-way		Binary				
	F_1	Acc.	Comp.	Cont.	Expa.	Expa.+EntRel	Temp.
P&C2012	-	-	31.32	49.82	-	79.22	26.57
J&E2015	-	-	35.93	52.78	-	80.02	27.63
Zhang2015	38.80	55.39	32.03	47.08	68.96	80.22	20.29
R&X2014	38.40	55.50	39.70	54.40	70.20	80.44	28.70
R&X2015	40.50	57.10	**41.00**	53.80	69.40	-	33.30
B&D2015	-	-	36.36	55.76	61.76	-	27.30
Liu2016	44.98	57.27	37.91	**55.88**	69.97	-	37.17
Ji2016	42.30	**59.50**	-	-	-	-	-
NNMA(two-level)	**46.29**	57.17	36.70	54.48	**70.43**	80.73	**38.84**
NNMA(three-level)	44.95	57.57	39.86	53.69	69.71	**80.86**	37.61

Table 4: Comparison with the State-of-the-art Approaches.

that the "Comparison" relation needs more passes of reading compared to the other three relations. The reason may be that the identification of the "Comparison" depends more on some deep analysis such as semantic parsing, according to (Zhou et al., 2010).

Next, we compare our models with six state-of-the-art baseline approaches, as shown in Table 4. The six baselines are introduced as follows.

- **P&C2012**: Park and Cardie (2012) designed a feature-based method and promoted the performance through optimizing the feature set.
- **J&E2015**: Ji and Eisenstein (2015) used two recursive neural networks on the syntactic parse tree to induce the representation of the arguments and the entity spans.
- **Zhang2015**: Zhang et al. (2015) proposed to use shallow convolutional neural networks to model two arguments respectively. We replicated their model since they used a different setting in preprocessing PDTB.
- **R&X2014, R&X2015**: Rutherford and Xue (2014) selected lexical features, production rules, and Brown cluster pairs, and fed them into a maximum entropy classifier. Rutherford and Xue (2015) further proposed to gather extra weakly labeled data based on the discourse connectives for the classifier.
- **B&D2015**: Braud and Denis (2015) combined several hand-crafted lexical features and word embeddings to train a max-entropy classifier.
- **Liu2016**: Liu et al. (2016) proposed to better classify the discourse relations by learning from other discourse-related tasks with a multi-task neural network.
- **Ji2016**: Ji et al. (2016) proposed a neural language model over sequences of words and used the discourse relations as latent variables to connect the adjacent sequences.

It is noted that P&C2012 and J&E2015 merged the "EntRel" relation into the "Expansion" relation[1]. For a comprehensive comparison, we also experiment our model by adding a *Expa.+EntRel vs Other* classification. Our NNMA model with two attention levels exhibits obvious advantages over the six baseline methods on the whole. It is worth noting that NNMA is even better than the R&X2015 approach which employs extra data.

As for the performance on each discourse relation, with respect to the F_1 measure, we can see that our NNMA model can achieve the best results on the "Expansion", "Expansion+EntRel" and "Temporal" relations and competitive results on the "Contingency" relation . The performance of recognizing the "Comparison" relation is only worse than R&X2014 and R&X2015. As (Rutherford and Xue, 2014) stated, the "Comparison" relation is closely related to the constituent parse feature of the text, like production rules. How to represent and

[1]EntRel is the entity-based coherence relation which is independent of implicit and explicit relations in PDTB. However some research merges it into the implicit Expansion relation.