			DF		NW	
Baseline			14.53		7.24	
Columbia/GWU			20.69		10.10	
Cornell-Pitt-Michigan			19.48		0.70	
Our work (best model)			**22.9**		**14.0**	
Target Type	Argument	Type/Subtype	DF-Relation	DF-Event	NW-Relation	NW-Event
Relation	Naive		7.7	17.3	5.1	9.7
Relation	Flat	Flat	11.1	19.2	7.2	9.7
Relation	Affine	Flat	15.9	21.5	8.2	13.0
Relation	Flat	Affine	17.3	21.7	**12.9**	12.4
Relation	Affine	Affine	22.0	**23.7**	12.8	**15.2**
Event	Naive		18.6	10.2	10.6	5.0
Event	Flat	Flat	17.9	13.0	9.5	7.4
Event	Affine	Flat	20.4	19.6	10.7	13.1
Event	Flat	Affine	**22.6**	22.4	12.8	14.8
Event	Affine	Affine	22.0	**23.7**	12.8	**15.2**

Table 1: Experimental results on the BeSt dataset. For our models, we analyze the performance for both types of targets by changing the encoding of specified target type. The other target type is encoded with affine maps throughout.

A Experimental Results

We provide a complete set of experimental results in Table 1. These results report the F1-score for both discussion forums and newswire for each target type. We also considered a *naive* encoding method for encoding relations and events which, similar to entity mention encoding, is the mean-pooling of the word representations corresponding to each relation and event span.