A Event-Relation Consistency Constraint

A pair of input tokens have positive temporal relation if and only if both tokens are events. This property is encoded in Table 8 with additional constraints that there can be only one label assigned to either token or relation. The following rules will satisfy this property.

1. $\sum_{P\in\mathcal{R}} r^P_{i,j} + r^N_{i,j} = 1$
2. $e^P_i + e^N_i = 1$
3. $e^P_i \geq r^P_{i,j}$ and $e^P_j \geq r^P_{i,j}$
4. $e^N_i + e^N_j \geq r^N_{i,j}$

Proof

$\rightarrow$) If either $e^P_i = 0$ or $e^P_i = 0$, then $r^P_{i,j}$ could only be 0. If $e^P_i = 1$ and $e^P_j = 1$, then $r^N_{i,j}$ can be either 0 or 1. Column $r^P_{i,j}$ is satisfied. If either $e^N_i = 1$ or / and $e^N_i = 1$, by Rule 4, $r^P_{i,j} = 0$ or 1. However, by Rule 2, one of the top three rows of Column e^P_i and e^P_j has to be true, which implies that $r^P_{i,j} = 0$ and thus, by Rule 1, $r^N_{i,j} = 1$. If $e^N_i = 0$ and $e^N_j = 0$, it's obvious that $r^N_{i,j}$ must be 0.

$\leftarrow$) If $r^P_{i,j} = 0$, then by Rule 3, e^P_i and e^P_j can be any number. If $r^P_{i,j} = 1$, then $e^P_i = 1$ and $e^P_j = 1$. If $r^N_{i,j} = 1$, by Rule 4, at least one of $e^P_i, e^P_j = 1$. If $r^N_{i,j} = 0$, it implies that $r^P_{i,j} = 1$ and hence $e^N_i = 1$ and $e^N_j = 1$ and therefore $e^P_i = 0$ and $e^P_j = 0$.

e^P_i	e^P_j	e^N_i	e^N_j	$r^P_{i,j}$	$r^N_{i,j}$
0	0	1	1	0	1
0	1	1	0	0	1
1	0	0	1	0	1
1	1	0	0	1, 0	0

Table 8: Event and Relation Global Constraint

B Evaluation Metrics Visualization

Each cell in Figure 4a and 4b is the count of predicted labels of gold pairs. In Figure 4a, S1 is the sum of column b, a, e, v, whereas in Figure 4b, S1 is the sum of b, a, e. Similar calculation applied to S1. The final Precision (P), Recall (R) and F1 scores are calculated as,

	TB-Dense		MATRES	
Single-task Model				
	Ent	Rel	Ent	Rel
hidden size	100	100	60	60
dropout	0.4	0.5	0.5	0.5
Multi-task Model				
hidden size	90		90	
dropout	0.6		0.3	
entity weight	6.0		16.0	
Pipeline Joint Model				
hidden size	90		90	
dropout	0.6		0.4	
entity weight	6.0		15.0	
Structured Joint Model				
lr	0.0005		0.001	
decay	0.1		0.1	
momentum	0.2		0.1	
$C_{\mathcal{E}}$	0.1		5.0	
T_{evt}	0.49		0.4	

Table 9: Best hyper-parameters.

	CogCompTime			**Pipeline Joint**			**Structured Joint**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
B	50.4	65.6	57.0	60.1	62.8	61.4	60.0	64.3	62.0
A	45.1	52.8	48.6	55.0	59.8	57.3	57.5	60.9	59.1
S	-	-	-	-	-	-	-	-	-
Avg	48.4	58.0	52.8	58.1	59.0	58.5	59.0	60.2	**59.6**

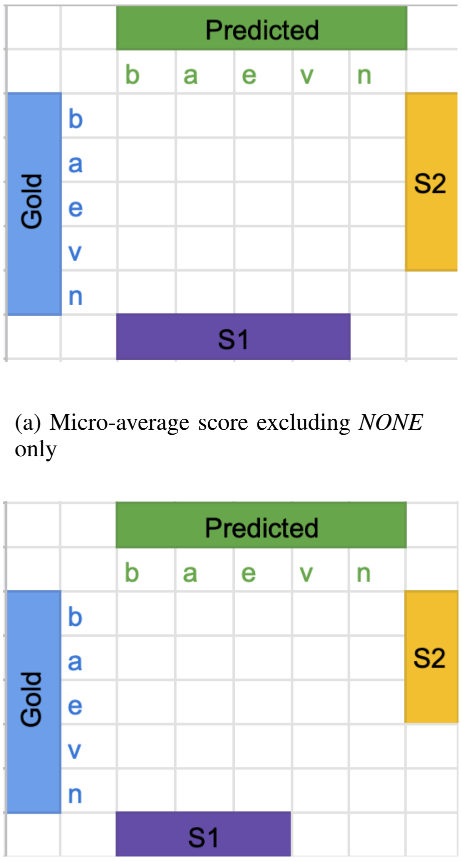
Table 10: Model Performance Breakdown for MATRES. *BEFORE* (**B**), *AFTER* (**A**), *SIMULTANEOUS* (**S**)

P = correct count / S1
R = correct count / S2
F1 = 2PR / (P+R)

C Best Hyper-Parameters

We observe that the Adam optimizer works well for single-task, multi-task and pipeline joint models, whereas the SGD optimizer works well for the structured joint model—possibly due to different loss functions used, i.e., cross-entropy loss vs. SSVM loss. We leave systematic investigation for future research. The best hyper-parameters can be found in Table 9.

D Performance Breakdown Tables



(a) Micro-average score excluding *NONE* only

(b) Micro-average score excluding *NONE* and *VAGUE*

Figure 4: Confusion matrix (table) with each cell representing count of predictions over each gold label. *BEFORE* (b); *AFTER* (a); *SIMULTANEOUS* (e); *VAGUE* (v); *NONE* (n).