Table 2: Dataset splits used for evaluation (indicated with ‡).

Split	Training data	#TLinks	#Documents	Test data	#TLinks	#Documents
TD‡	TD (train+dev)	4.4k	27	TD (test)	1.3k	9
TE3‡	TB, AQ, VC, TD (full)	17.5k	256	PT	0.9k	20

Table 3: Hyper-parameters from the experiments.

Hyper-parameter	Value
Document-creation starting time (s_{DCT})	0
Minimum event duration (d_{min})	0.1
Time-line margin (m_τ)	0.025
Hinge loss margin (m_h)	0.1
Dropout (α_d)	0.1
Word-level RNN units (α_{rnn})	25
Word-embedding size (α_{wemb})	50
POS-embedding size	10

Table 4: Evaluation of relative time-lines for each model and loss function, where L_* indicates the (unweighted) sum of L_τ, $L_{\tau ce}$, and $L_{\tau h}$.

	TE3‡			TD‡		
Model	**P**	**R**	**F**	**P**	**R**	**F**
Indirect: $O(n^2)$						
TL2RTL (L_τ)	53.5	51.1	52.3	59.1	61.2	60.1
TL2RTL ($L_{\tau ce}$)	53.9	51.7	**52.8**	61.2	60.7	60.9
TL2RTL ($L_{\tau h}$)	52.8	51.1	51.9	57.9	60.6	59.2
TL2RTL (L_*)	52.6	52.0	52.3	62.3	62.3	**62.3**
Direct: $O(n)$						
S-TLM (L_τ)	50.1	50.4	50.2	57.8	59.5	**58.6**
S-TLM ($L_{\tau ce}$)	50.1	50.0	50.1	53.4	53.5	53.5
S-TLM ($L_{\tau h}$)	51.5	51.7	**51.6**	55.1	56.4	55.7
S-TLM (L_*)	50.9	51.0	51.0	56.5	55.3	55.9
C-TLM (L_τ)	56.2	56.1	**56.1**	57.1	59.7	**58.4**
C-TLM ($L_{\tau ce}$)	54.4	55.4	54.9	52.4	57.3	54.7
C-TLM ($L_{\tau h}$)	55.7	55.5	55.6	55.3	54.9	55.1
C-TLM (L_*)	54.0	54.3	54.1	54.6	53.5	54.1

5 Results

We compared our three proposed models for the three loss functions L_τ, $L_{\tau ce}$, and $L_{\tau h}$, and their linear (unweighted) combination L_*, on TE3‡ and TD‡, for which the results are shown in Table 4.

A trend that can be observed is that overall performance on TD‡ is higher than that of TE3‡, even though less documents are used for training. We inspected why this is the case, and this is caused by a difference in class balance between both test sets. In TE3‡ there are many more TLinks of type *simultaneous* (12% versus 3%), which are very difficult to predict, resulting in lower scores for TE3‡ compared to TD‡. The difference in performance between the datasets is probably also be related to the dense annotation scheme of TD‡ compared to the sparser annotations of TE3‡, as dense annotations give a more complete temporal view of the training texts. For TL2RTL better TLink extraction[12] is also propagated into the final time-line quality.

If we compare loss functions L_τ, $L_{\tau ce}$, and $L_{\tau h}$, and combination L_*, it can be noticed that, although all loss functions seem to give fairly similar performance, L_τ gives the most robust results (never lowest), especially noticeable for the smaller dataset TD‡. This is convenient, because L_τ is fastest to compute during training, as it requires no score calculation for each TLink type. L_τ is also directly interpretable on the time-line. The combination of losses L_* shows mixed results, and has lower performance for S-TLM and C-TLM, but better performance for TL2RTL. However, it is slowest to compute, and less interpretable, as it is a combined loss.

Moreover, we can clearly see that on TE3‡, C-TLM performs better than the indirect models, across all loss functions. This is a very interesting result, as C-TLM is an order of complexity faster in prediction speed compared to the indirect models ($O(n)$ compared to $O(n^2)$ for a text with n entities).[13] We further explore why this is the case through our error analysis in the next section.

On TD‡, the indirect models seem to perform slightly better. We suspect that the reason for this is that C-TLM has more parameters (mostly the LSTM weights), and thus requires more data (TD‡ has much fewer documents than TE3‡) compared to the indirect methods. Another result supporting this hypothesis is the fact that the difference between C-TLM and S-TLM is small on the smaller

[12] F1 of 40.3 for TE3‡ and 48.5 for TD‡ (Ning et al., 2017)

[13] We do not directly compare prediction speed, as it would result in unfair evaluation because of implementation differences. However, currently, C-TLM predicts at ∼100 w/s incl. POS tagging, and ∼2000 w/s without. When not using POS, overall performance decreases consistently with 2-4 points.