A Supplementary Document

In this supplementary document, we provide more details about our dataset creation in Section B, model details in Section C and further experimental results in Section D.

B Data

In this Section, B.1 presents qualitative examples of our dataset, B.2 describe our overall data processing steps and the resulting changes in data composition per step, while B.3 explains our pre-processing steps per data source in detail.

B.1 Examples

A *Causal* example from each of the six corpora is presented in Table 6.

Corpus	*Causal* Example
AltLex	<ARG0>In the Philippines , Washi</ARG0> caused <ARG1>at least 1,268 deaths .</ARG1>
BECAUSE	<ARG0>Having only a Republican measure</ARG0> makes <ARG1>the task harder</ARG1>.
CTB	Iraq said it <ARG1>invaded</ARG1> Kuwait because of <ARG0>disputes</ARG0> over oil and money.
ESL	Ten <ARG1>dead</ARG1> in southern Iran <ARG0>quake</ARG0>.
PDTB	<ARG1>And the firms are stretching their nets far and wide</ARG1> <ARG0>to do it</ARG0>.
SemEval	The front <ARG0>wheels</ARG0> are making a <ARG1>grinding noise</ARG1> .

Table 6: *Causal* examples from each corpus. (<ARG0>, </ARG0>) marks the boundaries of a Cause span, while (<ARG1>, </ARG1>) marks the boundaries of a corresponding Effect span.

B.2 Data Processing – Overall

As mentioned in our main paper, we limited our study to examples of up to 3 sentences. For examples where a Cause and Effect argument spans within one sentence, we only retained the sentences containing the argument spans, even if they are not consecutive (i.e. "extended contexts" are not taken as inputs). This means that in most cases, each example comprises of 1 or 2 sentences. In some datasets (E.g. PDTB-3), the argument itself can span across multiple sentences. For such cases, the examples can comprise of up to 3 sentences. Table 7 reflects counts from the raw data, where the last column reflects the number of

examples we essentially drop from the final consolidated dataset. Subsequently, we perform train-test split, with the data distributions shown in Table 8. For datasets compatible with the Span Detection task, we also removed examples with more than 3 causal relations to simplify our current stage of modeling. The final data counts are available in Table 2.

Corpus	Number of Sentences				Total
	1	2	3	>3	
AltLex	1,027	-	-	-	1,027
BECAUSE	1,245	-	-	-	1,245
CTB	3,043	578	188	390	4,199
ESL	7,140	6,487	5,754	19,753	39,134
PDTB	28,088	22,945	1,468	1,175	53,676
SemEval	10,717	-	-	-	10,717
Total	51,260	30,010	7,410	21,318	109,998

Table 7: Sizes of raw datasets by sentence length.

Corpus	Split	Number of Sentences			Total
		1	2	3	
AltLex	Train	611	-	-	611
	Test	416	-	-	416
BECAUSE	Train	1,185	-	-	1,185
	Test	60	-	-	60
CTB	Train	2,671	496	150	3,317
	Test	372	82	38	492
ESL	Train	6,447	5,873	5,094	17,414
	Test	693	614	660	1,967
PDTB	Train	22,598	18,671	1,190	42,459
	Test	5,490	4,274	278	10,042
SemEval	Train	8,000	-	-	8,000
	Test	2,717	-	-	2,717
Total		51,260	30,010	7,410	88,680

Table 8: Sizes of train and test datasets by sentence length. We only retain examples with <=3 sentence length for our study.

B.3 Data Processing – Per Data Source

In this Section, we describe the processing steps done for each corpus. The codes are available under the '`processing`' folder in our repository.

Corpus	Split	Span
AltLex	Train	315
	Test	127
BECAUSE	Train	902
	Test	46
PDTB	Train	9,809
	Test	2,294
Total		13,493

Table 9: Sizes of final train and test datasets available for Span Detection task. These numbers are the ungrouped version corresponding to the grouped ones in Table 2.

AltLex [7] We used the development and gold datasets annotated by graduate students and crowd workers respectively. Although additional distantly labeled data are available in their repository, we wish to focus on a high-quality datasets and thus, incorporated only human-labelled datasets. Additionally, only the human-annotated datasets include labels with REASON and RESULT, which are needed to identify the Cause and Effect arguments. REASON and RESULT sentences were regarded as *Causal*, while sentences with the label NONE were treated as *Non-causal*. REASON examples were in the order of Effect-Signal-Cause, while RESULT had the format of Cause-Signal-Effect. Equipped with this template, we could convert the causal labels into Cause and Effect spans around the marked signal words. For train-test splits, the development set was used for testing, while the gold set was used for training. The AltLex corpus is suitable for all three tasks: For Span Detection, we had 415 examples; For sequence classification, we had 415 causal and 563 non-causal examples; For pair classification, we had access to 442 causal and 585 non-causal examples.

BECAUSE 2.0 [4] There are a few sentences where the Effect is annotated with no corresponding Cause[11] Such examples are treated as *Non-causal* in our consolidated corpus because we require both Cause and Effect spans to be present within each example for them to be considered *Causal*. On top of this treatment, BECAUSE 2.0 itself also contains some *Non-causal* examples. Additionally, we did not have access to the raw files of the NYT which requires subscription, and missed out on annotations for the 59 randomly selected articles. All in all, we obtained 965 *Causal* pairs and 280 *Non-causal* pairs. To fit our framework of creating a fixed subset for train-test split for easier benchmarking, we randomly set 10% of the documents as test sets.[12] For PTB files, we followed the

[11] An example sentence from BECAUSE 2.0 where the Effect is annotated with no corresponding cause is: *"However, the lower prices these retail chains are now expected to bring should make it easier for managers to raise the necessary capital and pay back the resulting <effect>debt</effect>."*.

[12] Test set: 'Article247_327.ann'

development and test splits recommended by PDTB V2.0 [14], mentioned later below.

CausalTimeBank (CTB) [9, 10] After filtering to examples with <= 3 sentence length, we ended up with 318 causal and 3,491 non-causal examples for pair classification, and 276 causal and 1,925 non-causal examples for sequence classification. To fit our framework of creating a fixed subset for train-test split for easier benchmarking, we randomly set 10% of the documents as test sets.[13] For files that tapped onto PTB, we followed the development and test splits recommended by PDTB V2.0 [14] which is described later below. Previous works [18, 10] evaluated models using 10-fold cross-validation or to use an external Temporal Eval-3 dataset [11]. Our method essentially only evaluates on one of the ten folds. Researchers can process the dataset to perform the other nine folds if needed. For Temporal Eval-3, we will work on including this dataset in the next release.

EventStoryLine (ESL) [2] We utilised the V1.0 expert annotations because it is the more updated version according to caselli-vossen-2017-event[14]. For events, we only consider markables tagged as actions or negative actions. All `PLOT_LINK` relations are treated as *Causal*.[15] To generate *Non-causal* examples, for every head event, we defined all tails unidentified by `PLOT_LINK` as *Non-causal*. Since ESL does not mark out the causal direction of their events, this dataset cannot be used for Pair Classification. ESL is only suitable for Sequence Classification, of which after filtering to keep only examples with <= 3 sentences, we obtained 2,232 examples. The last two topics were used as test set,[16] while the remaining 20 topics were used for training, as suggested by gao-etal-2019-modeling [5].

Penn Discourse Treebank V3.0 (PDTB) [16] The sense CONTINGENCY[17] was treated as a positive example of causal relations by previous researchers [13, 4]. Dunietz et al. [4] argued for the exclusion of evidentiary uses of causal language.[18] Thus, similar to the treatment by previous work [15], we treated the

[13] Test set: 'ea980120.1830.0456.tml', 'APW19980227.0494.tml', 'PRI19980306.2000.1675.tml', 'APW19980213.1320.tml', 'APW19980501.0480.tml', 'PRI19980205.2000.1998.tml'

[14] Compared to V0.9, V1.0 includes additional data curation checks to ensure consistency of TimeX3 normalisation, `TLINK` annotation and event annotation, plus removes some wrong event tokens.

[15] We ignore markables tagged as location, time or time date. TLINK relations that have time related entities (E.g. *"May 02"* or *"Thursday"*) are not difficult to classify as *Non-causal*. Hence, we excluded them from our corpus.

[16] Test set: T37, T41

[17] CONTINGENCY: one argument provides the reason, explanation or justification for the situation described by the other [16]

[18] E.g. *"We went to war based on bad intelligence."* [4]

two sub-senses that were similar to evidential types, BELIEF and SPEECHACT[19], as negative examples. For PDTB, we could retrace the Cause and Effect arguments in place based on the sub-sense. For five sub-senses,[20] the first argument is the Cause, while the second argument is the Effect. For four sub-senses,[21] the reverse. Examples with all other sub-senses were treated as *Non-causal*. The PDTB corpus is suitable for all three tasks. After filtering to examples with $<= 3$ sentences and $<= 3$ causal relations, we had 10,972 Span Detection, 41,669 Sequence Classification and 52,384 Pair Classification examples. For train-test split, we used the development and test set recommended by PDTB V2.0 [14],[22] while the remaining was used for training. Unless explicitly specified otherwise, we refer to the final test set as the combination of the development and test set.

SemEval 2010 Task 8 (SemEval) [6] We regarded the `Cause-Effect` relation as *Causal*, while all other relations were treated as *Non-causal*. Since the arguments annotated were short, SemEval is not compatible for Span Detection. Therefore, we only formatted SemEval for the Sequence Classification and Pair Classification task, of which, we obtained 10,690 and 10,717 examples respectively. We used their test set for testing, while all other examples were used for training.

C Model Details

For the BERT model, we used `bert-base-cased` from Huggingface [17]. Sequence output dimension from the BERT encoder is the default at 768. The token classifier had output dimensions of 5, while the sequence classifiers output dimension was 2. The classifiers' input dimension was the BERT embedding at 768. To train our model, we used the AdamW optimizer [8] with $\beta 1 = 0.9$, $\beta 2 = 0.999$. Learning rate was set at $2e-05$ with linear decay. GPU train batch size was 128. Maximum sequence length was variable according to the longest sequence per batch. 20 epochs were ran per experiment.

Following previous works [1, 12, 3], we repeated each experiment 5 times with different random seeds.[23] We reported mean and standard deviation performance in the results section. We also performed Paired T-test when comparing models with one another (E.g. UniCausal ($\alpha = 1$) *vs.* Individual) to calculate statistical significance of difference in means.

[19] BELIEF: One argument gives the evidence justifying the claim in the other argument; SPEECHACT: One argument provides the reason for the speaker uttering the speech act in the other argument [16]

[20] The first argument is Cause span for: CONTINGENCY.CAUSE.RESULT, CONTINGENCY.PURPOSE.ARG1-AS-GOAL, CONTINGENCY.CONDITION.ARG1-AS-COND, CONTINGENCY.NEGATIVE-CONDITION.ARG1-AS-NEGCOND, and CONTINGENCY.NEGATIVE-CAUSE.NEGRESULT

[21] The second argument is Cause span for: CONTINGENCY.CAUSE.REASON, CONTINGENCY.PURPOSE.ARG2-AS-GOAL, CONTINGENCY.CONDITION.ARG2-AS-COND, and CONTINGENCY.NEGATIVE-CONDITION.ARG2-AS-NEGCOND

[22] Dev set: wsj_00, wsj_01, wsj_24 ; Test set: wsj_22, wsj_23

[23] The random seeds used were: 42, 975, 106, 239, and 303.

For a complete run to train and test on all six datasets, the Individual models for Sequence Classification, Span Detection, and Pair classification took $1h6m38s$, $34m21s$, and $1h1m27s$ to run in total. In terms of memory, each model requires approximately $20000MiB$ GPU space. In terms of storage, each saved model is about $420MB$ in size.

D Experiments

D.1 Performance Across Epochs

Figure 2 presented the F1 scores of both each of the Individual models for each of the three tasks across training epochs. Span Detection observed the largest F1 score improvements in the first few epochs. All three tasks generally plateaued in their F1 improvements towards the second half of training.

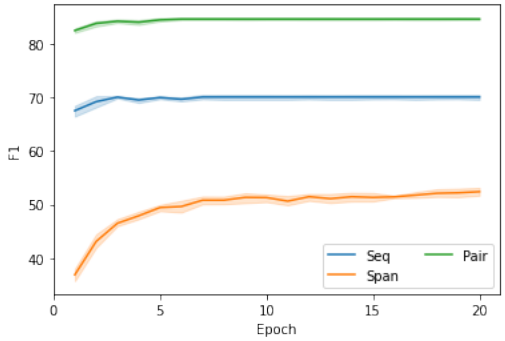

Fig. 2: F1 scores for all datasets across three models over training epochs: Sequence Classification Binary F1 in Blue, Pair Classification Binary F1 in Green, and Span Detection Macro F1 in Orange. The shaded regions reflect confidence intervals.

D.2 Additional Results

		(I) Sequence Classification				(II) Span Detection			(III) Pair Classification			
Test Set	Training Set	P	R	F1	Acc	P	R	F1	P	R	F1	Acc
All	All	**71.13**	69.14	**70.10**	**86.27**	**46.33**	**60.35**	**52.42**	**85.44**	83.93	**84.68**	**93.68**
		±0.80	±1.60	**±0.58**	**±0.15**	**±1.22**	**±0.30**	**±0.90**	**±0.96**	±0.44	**±0.27**	**±0.16**
	AltLex	39.77	29.03	32.93	72.90	4.29	11.21	6.20	31.68	44.64	31.83	61.94
		±3.17***	±7.03***	±3.57***	±2.41***	±0.59***	±1.01***	±0.74***	±9.75***	±23.97*	±3.93***	±17.68*
	BECAUSE	24.74	**94.25**	39.15	31.62	10.35	16.57	12.74	22.64	**93.11**	36.40	32.27
		±1.02***	**±4.07*****	±0.99***	±5.62***	±0.29***	±0.59***	±0.35***	±0.70***	**±3.84*****	±0.64***	±4.28***
	CTB	66.88	22.66	33.49	79.30	-	-	-	73.66	11.83	20.17	80.78
		±3.05*	±4.97***	±5.48***	±0.36***				±6.50*	±3.99***	±5.78***	±0.61***
	ESL	28.48	65.37	39.62	53.52	-	-	-	-	-	-	-
		±1.32***	±3.09*	±0.89***	±3.61***							
	PDTB	63.30	58.85	60.99	82.47	45.88	59.92	51.97	59.46	79.86	68.13	84.45
		±0.66***	±1.29***	±0.76***	±0.26***	±0.65	±0.27***	±0.48	±2.15***	±1.46***	±0.88***	±0.88***
	SemEval	52.00	19.44	28.25	77.02	-	-	-	61.08	17.23	26.66	80.37
		±2.64***	±1.12***	±0.86***	±0.44***				±7.17***	±2.11***	±1.86***	±0.65***
AltLex	All	**50.76**	63.48	**56.37**	71.87	**27.74**	**42.99**	**33.72**	82.60	87.09	**84.76**	**90.43**
		±1.61	±4.60	**±2.49**	±1.19	**±1.20**	**±0.85**	**±1.12**	±1.99	±1.53	**±0.66**	**±0.55**
	AltLex	50.58	53.57	51.85	71.52	16.06	32.28	21.45	**88.70**	74.17	80.57	89.04
		±3.35	±5.17*	±2.53	±1.99	±1.45***	±2.71***	±1.87***	**±7.47**	±3.40***	±2.48*	±1.80
	BECAUSE	31.69	91.83	47.02	40.40	5.65	10.71	7.38	32.14	**94.80**	47.99	37.21
		±1.90***	±5.45***	±1.52***	±7.17***	±1.89***	±2.42***	±2.19***	±1.18***	**±1.98*****	±1.33***	±3.34***
	CTB	63.35	51.65	55.91	**77.36**	-	-	-	84.28	12.60	19.16	71.83
		±3.23***	±13.15	±7.63	**±1.13****				±18.29	±13.35***	±5.64***	±1.64***
	ESL	30.83	**93.57**	46.29	37.66	-	-	-	-	-	-	-
		±1.37***	**±6.81*****	±1.15***	±5.74***							
	PDTB	48.02	49.91	48.94	70.12	5.36	9.06	6.73	26.73	82.20	40.34	25.82
		±2.10	±2.00***	±1.88***	±1.28	±0.81***	±1.11***	±0.94***	±0.89***	±4.26	±1.52***	±1.14***
	SemEval	55.23	19.65	28.95	72.37	-	-	-	72.07	25.35	37.07	74.13
		±2.44*	±1.58***	±1.74***	±0.48				±2.90***	±6.21***	±6.58***	±0.91***

BEC-AUSE	All	92.32 ±1.69	70.24 ±2.04	79.77 ±1.68	71.37 ±2.24	**32.51 ±2.82**	44.30 ±2.33	37.47 ±2.57	**87.93 ±1.73**	94.78 ±1.94	**91.21 ±1.18**	**86.00 ±1.90**
	AltLex	98.67 ±2.98*	22.93 ±8.73**	36.47 ±11.18**	37.65 ±6.41**	8.30 ±1.40**	18.92 ±2.10**	11.51 ±1.63**	80.49 ±7.19	37.83 ±21.33*	48.44 ±20.00*	45.00 ±12.19**
	BECAUSE	86.20 ±3.31**	96.10 ±5.06**	**90.77 ±2.22***	**84.31 ±3.67***	31.15 ±5.53	**48.17 ±5.66**	**37.79 ±5.77**	83.37 ±2.33*	**97.83 ±2.17**	90.01 ±1.95	83.33 ±3.33
	CTB	**97.87 ±2.94***	38.54 ±8.86**	54.73 ±9.40**	49.80 ±6.59**	-	-	-	84.90 ±11.92	13.04 ±7.37**	22.00 ±10.92**	31.67 ±4.86***
	ESL	83.42 ±1.02**	**98.05 ±2.67***	90.12 ±1.05**	82.75 ±1.64**	-	-	-	-	-	-	-
	PDTB	89.33 ±1.57*	57.07 ±2.78**	69.61 ±2.16**	60.00 ±2.24**	32.50 ±2.77	40.00 ±1.92	35.84 ±2.42	77.99 ±1.47**	87.83 ±3.95**	82.59 ±2.17**	71.67 ±3.12***
	SemEval	100.00 ±0.00**	9.27 ±2.04**	16.91 ±3.40**	27.06 ±1.64***	-	-	-	98.18 ±4.07**	15.22 ±7.37**	25.70 ±11.46**	34.67 ±5.32***
CTB	All	42.37 ±2.11	66.19 ±4.26	51.58 ±1.82	83.48 ±1.21	-	-	-	75.66 ±3.61	72.50 ±6.81	**73.94 ±4.68**	95.04 ±0.78
	AltLex	40.43 ±6.72	39.05 ±13.94*	38.21 ±6.20*	83.80 ±2.37	-	-	-	20.48 ±9.28**	35.42 ±24.87*	20.06 ±7.14**	74.84 ±14.26*
	BECAUSE	14.45 ±0.92**	**98.10 ±1.99***	25.17 ±1.34**	22.15 ±7.02***	-	-	-	13.42 ±0.99**	**97.50 ±0.93***	23.58 ±1.52**	38.09 ±5.14***
	CTB	**71.46 ±4.90***	58.57 ±0.59	**63.65 ±5.55***	**91.27 ±0.62*****	-	-	-	**82.18 ±7.47**	67.08 ±9.93	73.29 ±6.14	**95.28 ±0.96**
	ESL	18.48 ±0.91**	93.33 ±4.58**	30.84 ±1.35**	44.30 ±3.50***	-	-	-	-	-	-	-
	PDTB	33.36 ±1.77**	48.57 ±2.71**	39.54 ±1.88**	80.25 ±0.86*	-	-	-	16.97 ±1.26**	64.17 ±13.13	26.74 ±2.42**	65.98 ±3.86***
	SemEval	49.19 ±4.13*	31.90 ±4.33**	38.51 ±3.44**	86.52 ±0.76*	-	-	-	46.94 ±3.96**	55.42 ±3.49**	50.63 ±1.74**	89.43 ±1.09***
ESL	All	76.11 ±2.04	67.43 ±3.45	71.45 ±1.89	73.79 ±1.34	-	-	-	-	-	-	-

	AltLex	45.40 ±3.38**	65.84 ±18.33	53.30 ±8.37**	45.60 ±3.77***	-	-	-	-	-	-	-
	BECAUSE	48.17 ±1.05**	**93.27 ±5.89***	63.49 ±1.94**	47.84 ±1.95***	-	-	-	-	-	-	-
	CTB	**92.10 ±7.32***	21.42 ±12.01**	33.26 ±15.44**	60.69 ±4.70**	-	-	-	-	-	-	-
	ESL	75.90 ±5.08	87.79 ±5.06**	**81.21 ±2.35***	**80.17 ±3.00****	-	-	-	-	-	-	-
	PDTB	68.44 ±8.81	27.26 ±3.39**	38.71 ±3.15**	58.10 ±2.16***	-	-	-	-	-	-	-
	SemEval	42.91 ±1.95**	49.56 ±5.70**	45.95 ±3.50**	43.45 ±1.86***	-	-	-	-	-	-	-
PDTB	All	72.59 ±0.61	66.34 ±1.63	69.31 ±0.70	84.63 ±0.17	**47.77 ±1.22**	**61.54 ±0.29**	**53.78 ±0.88**	84.56 ±1.17	82.04 ±0.46	83.28 ±0.36	92.43 ±0.23
	AltLex	34.38 ±2.46**	17.92 ±6.86**	22.91 ±5.79**	69.49 ±1.84**	3.77 ±0.60**	9.97 ±1.04**	5.47 ±0.76**	24.08 ±3.84**	39.15 ±28.09*	25.11 ±8.75**	54.18 ±20.53*
	BECAUSE	27.49 ±0.92**	**93.87 ±4.32*****	42.49 ±0.68**	33.39 ±4.95**	10.28 ±0.28**	16.28 ±0.58**	12.60 ±0.34**	24.24 ±0.52**	**92.47 ±4.46*****	38.39 ±0.37**	31.81 ±3.78***
	CTB	66.42 ±3.44**	16.30 ±3.22**	25.97 ±3.73**	75.87 ±0.09***	-	-	-	83.73 ±8.27	3.77 ±3.43**	7.02 ±6.06**	77.74 ±0.65***
	ESL	33.97 ±1.57**	57.18 ±3.51**	42.55 ±1.25**	59.58 ±2.58***	-	-	-	-	-	-	-
	PDTB	**73.54 ±0.37***	67.35 ±0.90	**70.31 ±0.56***	**85.11 ±0.22*****	48.89 ±0.54	62.92 ±0.27**	55.02 ±0.38*	**86.28 ±0.38**	81.27 ±0.36**	**83.70 ±0.34**	**92.73 ±0.15**
	SemEval	27.47 ±0.84**	6.24 ±1.24**	10.11 ±1.61**	71.13 ±0.65***	-	-	-	29.32 ±3.92**	4.80 ±2.19**	8.08 ±3.20**	75.46 ±0.86***
Sem-Eval	All	73.39 ±1.18	89.51 ±1.59	80.64 ±0.46	94.81 ±0.16	-	-	-	93.38 ±0.88	**96.10 ±0.59**	94.71 ±0.23	98.70 ±0.07
	AltLex	44.85 ±9.44**	76.89 ±5.62*	55.83 ±6.68**	84.68 ±5.49*	-	-	-	50.83 ±19.73*	73.90 ±5.42**	57.72 ±14.52*	84.33 ±12.18

BECAUSE	13.52 ±1.28**	**97.13 ±2.20***	23.71 ±1.93**	23.91 ±8.67***	-	-	-	14.56 ±1.40**	95.61 ±2.53	25.23 ±2.02**	31.00 ±8.75***
CTB	63.55 ±4.28**	46.16 ±16.86*	51.76 ±13.85*	90.27 ±1.03***	-	-	-	68.73 ±7.25**	59.63 ±6.21**	63.69 ±5.65**	91.79 ±1.36***
ESL	15.33 ±1.93**	91.40 ±7.82	26.15 ±2.62**	36.57 ±12.82***	-	-	-				
PDTB	18.60 ±2.29**	21.16 ±4.87**	19.75 ±3.35**	79.40 ±0.51***	-	-	-	22.25 ±1.33**	70.24 ±9.89**	33.64 ±1.76**	66.60 ±4.06***
SemEval	**87.84 ±1.49***	91.40 ±0.26*	**89.58 ±0.71***	**97.43 ±0.20*****	-	-	-	**93.96 ±0.49**	95.67 ±0.59	**94.80 ±0.28**	**98.73 ±0.07**

Table 10: Performance metrics for different test datasets across the three tasks using Individual models. The top score for each metric per test set is bolded. Tasks that are not applicable to the dataset are indicated by "-". Scores are reported in percentages (%). Paired T-test of the models was conducted against setting where all datasets were used for training, with statistical significance indicated by: ***< 0.001, **< 0.01, *< 0.05. For each test set, the first row (i.e. when training set is "All") is consolidated and corresponds to Table 3. The F1 scores of this table also corresponds to the values consolidated in Table 5.

References

1. Akbik, A., Bergmann, T., Vollgraf, R.: Pooled contextualized embeddings for named entity recognition. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 724–728. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1078, https://aclanthology.org/N19-1078
2. Caselli, T., Vossen, P.: The event StoryLine corpus: A new benchmark for causal and temporal relation extraction. In: Proceedings of the Events and Stories in the News Workshop. pp. 77–86. Association for Computational Linguistics, Vancouver, Canada (Aug 2017). https://doi.org/10.18653/v1/W17-2711, https://aclanthology.org/W17-2711
3. Chiu, J.P., Nichols, E.: Named entity recognition with bidirectional LSTM-CNNs. Transactions of the Association for Computational Linguistics **4**, 357–370 (2016). https://doi.org/10.1162/tacl_a_00104, https://aclanthology.org/Q16-1026
4. Dunietz, J., Levin, L., Carbonell, J.: The BECauSE corpus 2.0: Annotating causality and overlapping relations. In: Proceedings of the 11th Linguistic Annotation Workshop. pp. 95–104. Association for Computational Linguistics, Valencia, Spain (Apr 2017). https://doi.org/10.18653/v1/W17-0812, https://aclanthology.org/W17-0812
5. Gao, L., Choubey, P.K., Huang, R.: Modeling document-level causal structures for event causal relation identification. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 1808–1817. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1179, https://aclanthology.org/N19-1179
6. Hendrickx, I., Kim, S.N., Kozareva, Z., Nakov, P., Ó Séaghdha, D., Padó, S., Pennacchiotti, M., Romano, L., Szpakowicz, S.: SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In: Proceedings of the 5th International Workshop on Semantic Evaluation. pp. 33–38. Association for Computational Linguistics, Uppsala, Sweden (Jul 2010), https://aclanthology.org/S10-1006
7. Hidey, C., McKeown, K.: Identifying causal relations using parallel Wikipedia articles. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1424–1433. Association for Computational Linguistics, Berlin, Germany (Aug 2016). https://doi.org/10.18653/v1/P16-1135, https://aclanthology.org/P16-1135
8. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net (2019), https://openreview.net/forum?id=Bkg6RiCqY7
9. Mirza, P., Sprugnoli, R., Tonelli, S., Speranza, M.: Annotating causality in the TempEval-3 corpus. In: Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL). pp. 10–19. Association for Computational Linguistics, Gothenburg, Sweden (Apr 2014). https://doi.org/10.3115/v1/W14-0702, https://aclanthology.org/W14-0702
10. Mirza, P., Tonelli, S.: An analysis of causality between events and its relation to temporal information. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. pp. 2097–2106. Dublin

City University and Association for Computational Linguistics, Dublin, Ireland (Aug 2014), https://aclanthology.org/C14-1198
11. Mirza, P., Tonelli, S.: CATENA: CAusal and TEmporal relation extraction from NAtural language texts. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. pp. 64–75. The COLING 2016 Organizing Committee, Osaka, Japan (Dec 2016), https://aclanthology.org/C16-1007
12. Peters, M.E., Ammar, W., Bhagavatula, C., Power, R.: Semi-supervised sequence tagging with bidirectional language models. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1756–1765. Association for Computational Linguistics, Vancouver, Canada (Jul 2017). https://doi.org/10.18653/v1/P17-1161, https://aclanthology.org/P17-1161
13. Ponti, E.M., Korhonen, A.: Event-related features in feedforward neural networks contribute to identifying causal relations in discourse. In: Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics. pp. 25–30. Association for Computational Linguistics, Valencia, Spain (Apr 2017). https://doi.org/10.18653/v1/W17-0903, https://aclanthology.org/W17-0903
14. Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A.K., Webber, B.L.: The penn discourse treebank 2.0. In: Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco. European Language Resources Association (2008), http://www.lrec-conf.org/proceedings/lrec2008/summaries/754.html
15. Tan, F.A., Hürriyetoğlu, A., Caselli, T., Oostdijk, N., Nomoto, T., Hettiarachchi, H., Ameer, I., Uca, O., Liza, F.F., Hu, T.: The causal news corpus: Annotating causal relations in event sentences from news. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. pp. 2298–2310. European Language Resources Association, Marseille, France (Jun 2022), https://aclanthology.org/2022.lrec-1.246
16. Webber, B., Prasad, R., Lee, A., Joshi, A.: The penn discourse treebank 3.0 annotation manual. Philadelphia, University of Pennsylvania (2019)
17. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45. Association for Computational Linguistics, Online (Oct 2020). https://doi.org/10.18653/v1/2020.emnlp-demos.6, https://aclanthology.org/2020.emnlp-demos.6
18. Zuo, X., Cao, P., Chen, Y., Liu, K., Zhao, J., Peng, W., Chen, Y.: Improving event causality identification via self-supervised representation learning on external causal statement. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 2162–2172. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.findings-acl.190, https://aclanthology.org/2021.findings-acl.190