A Counts of discourse relations on Models

Here, we show the model we have trained: An lstm on labeled data (we call it *lbl-lstm*), and its fine tuned version on pseudo labeled data (let us call it *ft-lstm*), and a version obtained by the second iteration of self-training (call it *ft-itr2-lstm*). On the reverse side, we applied combination of reverse model reranking and self-training on 1, 2 and 3 iterations (we call these models *ft-rrk-rev-lstm ft-rrk-rev-itr2-lstm* and *ft-rrk-rev-itr3-lstm* respectively). The tables (see below) show how this models behave on challenging and standard test sets.[6]

- The challenge test has: Unlike - 80; Like - 80; No Rel - 77.
- The standard test has: Unlike - 166; like - 495; No Rel - 138.

We classify errors made by a model with respect to rhetorical relations, referring to them as follows: Unlike & Like - if model generates *like* (i.e. SIMILARITY) instead of *unlike* (i.e. CONTRAST); Like & Unlike - if model generates *unlike* (i.e. CONTRAST) instead of *like* (i.e. SIMILARITY); Rest - stand for the cases where model makes errors with respect to generating either *unlike* (i.e. CONTRAST) or *like* (i.e. SIMILARITY) that do not fall in to the cases described above (e.g. generates unlike where there is neither SIMILARITY nor CONTRAST in the content plans).

Model	Unlike	Like	UnlikeVSLike	LikeVSUnlike	NoRel	Rest
ft-itr2	58	72	15	7	76	9
ft-lstm	65	72	9	6	77	8
ft-rrk-rev-itr2	64	72	9	0	76	18
ft-rrk-rev-itr3	71	77	5	2	74	8
ft-rrk-rev	66	75	7	4	76	9
lbl-lstm	72	72	4	6	77	6

Table 4: Challenge Test: Fact-SM

Model	Unlike	Like	UnlikeVSLike	LikeVSUnlike	NoRel	Rest
ft-itr2	146	454	20	41	136	2
ft-lstm	140	460	25	35	138	1
ft-rrk-rev-itr2	162	493	4	1	131	8
ft-rrk-rev-itr3	161	494	5	1	121	17
ft-rrk-rev	149	462	17	33	133	5
lbl-lstm	149	456	17	39	133	5

Table 5: Standard Test: Fact-SM

Model	Unlike	Like	UnlikeVSLike	LikeVSUnlike	NoRel	Rest
ft-itr2	166	495	0	0	138	0
ft-lstm	166	495	0	0	135	3
ft-rrk-rev-itr2	166	495	0	0	138	0
ft-rrk-rev-itr3	166	495	0	0	138	0
ft-rrk-rev	166	495	0	0	138	0
lbl-lstm	166	495	0	0	134	4

Table 6: Standard Test: RST-SM

Model	Unlike	Like	UnlikeVSLike	LikeVSUnlike	NoRel	Rest
ft-itr2	74	80	0	0	77	6
ft-lstm	75	79	0	1	77	5
ft-rrk-rev-itr2	68	78	0	0	77	14
ft-rrk-rev-itr3	71	78	0	0	77	11
ft-rrk-rev	73	79	0	0	77	8
lbl-lstm	73	80	0	0	73	11

Table 7: Challenge Test: RST-SM

Model	Unlike	Like	UnlikeVSLike	LikeVSUnlike	NoRel	Rest
ft-itr2	68	80	0	0	77	12
ft-lstm	79	80	0	0	75	3
ft-rrk-rev-itr2	74	80	0	0	77	6
ft-rrk-rev-itr3	67	79	0	0	77	14
ft-rrk-rev	76	79	0	0	77	5
lbl-lstm	72	80	0	0	75	10

Table 8: Challenge Test: RST-LG

Model	Unlike	Like	UnlikeVSLike	LikeVSUnlike	NoRel	Rest
ft-itr2	151	495	0	0	138	15
ft-lstm	166	495	0	0	136	2
ft-rrk-rev-itr2	164	494	0	0	137	4
ft-rrk-rev-itr3	162	494	0	0	138	5
ft-rrk-rev-lstm	166	495	0	0	134	4
lbl-lstm	166	495	0	0	131	7

Table 9: Standard Test: RST-LG

Model	Unlike	Like	UnlikeVSLike	LikeVSUnlike	NoRel	Rest
ft-itr2	66	74	5	2	77	13
ft-lstm	64	75	7	1	77	13
ft-rrk-rev-itr2	71	74	0	2	77	13
ft-rrk-rev-itr3	71	75	0	1	77	13
ft-rrk-rev	71	74	0	2	77	13
lbl-lstm	69	76	3	1	77	11

Table 10: Challenge Test: FACT-LG

Model	Unlike	Like	UnlikeVSLike	LikeVSUnlike	NoRel	Rest
ft-itr2	157	492	9	3	138	0
ft-lstm	147	465	16	30	138	3
ft-rrk-rev-itr2	164	493	2	2	138	0
ft-rrk-rev-itr3	166	493	0	2	138	0
ft-rrk-rev	158	493	8	2	138	0
lbl-lstm	142	463	24	32	137	1

Table 11: Standard Test: FACT-LG

[6] On small models, the numbers are average, each model trained twice.

B Reproducibility Details

Datasets We conduct experiments on the RST-Large and RST-Small datasets, which have their FACT versions (obtained by erasing of rhetorical relations from the content plans). The task consists of 4304 parallel items for training, and 422 for validation. There are 136 unique tokens in the content plans (meaning representations), and 160 in the texts (which are delexicalized and anonymized texts of English).

Implementation Our implementation of self-training and reverse model reranking is based on one-layer LSTM with attention. We use the open source `fairseq` implementation (Ott et al., 2019). The word embedding and hidden size dimensions are 300 and 128 respectively, and the decoder output embedding size is 512. The dropout rate for both encoder and decoder is 0.2. There are no more than 128 sentences in a batch. Training uses early stopping when the validation loss has not improved for the last 20 epochs. The learning rate is 0.001, and the scheduler is *ReduceLROnPlateau* whose factor is 0.1 and patience is 3. The maximum output length is 2 times source length plus 50, and the beam size is 5. The loss function is optimized with Adam (Kingma and Ba, 2014), where $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$.

Training Details For every experiment, the computing infrastructure we used is an NVIDIA V100 GPU and an Intel(R) Xeon(R) Platinum 8268 CPU @ 2.90GHz CPU. We below provide details for RST models, as Fact-Smalls do not show significance difference. The numbers of (trainable) parameters of models for RST-Small and RST-Large datasets are 800840. Training a lstm model on the RST dataset (i.e. with no pseudo labeled data) takes around 2k seconds for 60 epochs. Training a model on the pseudo-labeled 80743 dataset takes around 30K seconds for 57 epochs. Training and validation loss at convergence is around 1.32. As mentioned previously, the speed of decoding was 37,973 tokens/s. Except the vanilla decoding, the speed of reverse model reranking was 15,1681.26 tokens/s.

C Model performances on Challenge and Standard test sets

Names of the models are as follows: An lstm on labeled data (we call it *lbl-lstm*), and its fine tuned version on pseudo labeled data (*ft-lstm*), and a version obtained by the second iteration of self-training (*ft-itr2-lstm*). On the reverse side, the combinations of reverse model reranking and self-training on 1, 2 and 3 iterations are refereed to as, *ft-rrk-rev-lstm ft-rrk-rev-itr2-lstm* and *ft-rrk-rev-itr3-lstm*, respectively.

Here we show the errors of the models with respect to reparations, omissions and hallucinations (ROHs) as follows: For a given test example, if a model makes reparations, and/or makes omissions and/or hallucinations, we count 1 error on that example. In case a model produces neither reparations, nor omissions, nor hallucinations, we count no error.

Figures 6 and 7 show the best performing models with respect to minimizing ROH errors.

It must be underlined that the models are unstable with respect to ROH errors though and different hyper parameters may produce even statistically different results with respect to ROHs. Nevertheless, the these figures show that at their best performances RST-SM performs better than others.

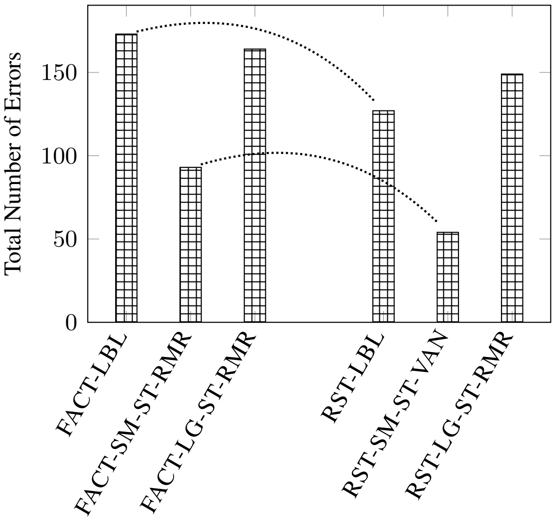

Figure 6: The challenge test set: ROH Errors by model. Significant differences between their performance in terms of Fisher's Exact Test statistics (we take the significance threshold 5%) are marked by a link between the corresponding models.

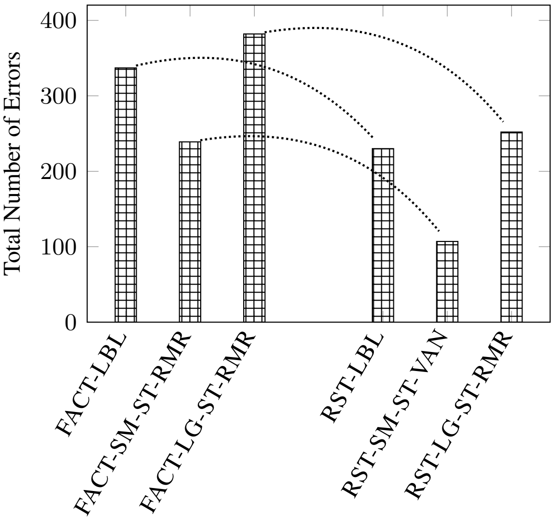

Figure 7: The standard test set: ROH errors by model. Significant differences between their performance in terms of Fisher's Exact Test statistics (we take the significance threshold 5%) are marked by a link between the corresponding models.

D BLEU4 Scores

lbl-lstm	53.51
ft-lstm	69.12
ft-itr2-lstm	54.34
hufft-rrk-rev-lstm	64.585
ft-rrk-rev-itr2-lstm	67.64
ft-rrk-rev-itr3-lstm	61.915

Table 12: FACT-SM on Challenge

lbl-lstm	71.99
ft-lstm	73.1
ft-itr2-lstm	74.88
ft-rrk-rev-lstm	78.705
ft-rrk-rev-itr2-lstm	78.04
ft-rrk-rev-itr3-lstm	79.545

Table 13: FACT-SM on Standard

lbl-lstm	84.385
ft-lstm	86.855
ft-itr2-lstm	84.01
ft-rrk-rev-itr2	86.055
ft-rrk-rev-itr3-lstm	86.415
ft-rrk-rev-lstm	86.74

Table 14: RST-Small on Standard

lbl-lstm	72.85
ft-lstm	74.53
ft-itr2-lstm	76.18
ft-rrk-rev-lstm	74.90
ft-rrk-rev-itr2-lstm	77.00
ft-rrk-rev-itr3-lstm	78.61

Table 15: RST-Small on Challenge

lbl-lstm	83.79
ft-lstm	81.62
ft-itr2-lstm	81.75
ft-rrk-rev-lstm	82.62
ft-rrk-rev-itr2	84.33
ft-rrk-rev-itr3-lstm	83.88

Table 16: RST-Large on Standard

lbl-lstm	60.81
ft-lstm	64.77
ft-itr2-lstm	66.83
ft-rrk-rev-lstm	6 69.50
ft-rrk-rev-itr2	70.53
ft-rrk-rev-itr3-lstm	74.17

Table 17: RST-Large on Challenge

lbl-lstm	73.7
ft-lstm	75.985
ft-itr2-lstm	75.3
ft-rrk-rev-lstm	76.33
ft-rrk-rev-itr2	79.795
ft-rrk-rev-itr3-lstm	81.435

Table 18: FACT-Large on Standard

lbl-lstm	54.78
ft-lstm	58.36
ft-itr2-lstm	56.13
ft-rrk-rev-lstm	57.85
ft-rrk-rev-itr	61.25
ft-rrk-rev-itr3-lstm	63.19

Table 19: FACT-Large on Challenge

E Hallucination, Repetition, and Omission

Even the RST-SM-ST-VAN model, which otherwise performs very well, produces both hallucinations (of the exhibit item's creation time) and repetitions (of the period in which the exhibit item is created).

T this is a figurine and it was created during historical-period0 . like the stamnos you recently saw , this figurine is made of clay-material0 .

H this is a figurine and it was created during historical-period0 . like the stamnos you recently saw , this figurine is made of clay-material0 . it was created in entity0-creation-time during historical-period0 .

Notice both this error and the error shown in section 8.1 occur when the target is short. While we don't conclude length is the definitive source of such errors, it seems the models expected the content in these items to be longer than it is.

The following error from RST-SM-ST-VAN shows omission of the exhibit item's painting technique.

T this is a rhyton ; it was made by potter0 and it was originally from region0 . like the other exhibits you recently saw , this rhyton was created during historical-period0 . it was created in entity0-creation-time ; it was painted with painting-technique0 and it entity0-exhibit-form .

H this is a rhyton ; it was made by potter0 and it was originally from region0 . like the other exhibits you recently saw , this rhyton was created during historical-period0 . it was created in entity0-creation-time and it entity0-exhibit-form .

The painting technique property is frequent, yet is mysteriously omitted here.