A Preprocessing of back-translations

We aligned the back-translated data with original InScript stories in terms of events and participants. The alignment relies on string and position matching as well as synonyms and similar words to map the event and participant labels from the original to the paraphrases. If the labeled source token/its lemma appears in the same position in the original as in the back-translated story, we copy the label.

We also consider similar words regardless of the position. We search for paraphrases using WordNet (Fellbaum, 1998) by considering the synonyms of labeled source words. For each labeled source word, we also consider 10 closest vectors as found in the embedding space of word2vec-google-news-300 pretrained vectors (Řehůřek and Sojka, 2010) to see if any of these neighboring words appear in the back-translated story.

We were not able to map all labels from the original to the back-translation: the latter contain 18% fewer event, and 10% fewer participant labels. The BLEU score between the original and back-translation is 58.22.

B Implementation Details

$80\%, 10\%, 10\%$ of the stories are randomly selected and designated as the training, validation and test set, respectively. The model is implemented with AllenNLP 1.0 (Gardner et al., 2017) . To regularize the model, dropout (Srivastava et al., 2014) with a universal rate is applied to all dense layers, in conjunction with weight decay. We use gradient norm clipping to stabilize the training. The optimization is performed with adam (Kingma and Ba, 2014) in conjunction with early-stopping which monitors validation loss, and the hyper-parameter tuning is performed with random hyper-parameter search (Bergstra and Bengio, 2012). Optimization takes on average 3.5 ours on a singe Tesla v100. We performed 20 trial for choosing the hyper-parameters, and 5 parallel optimizations in order to do the significance test.

The hyper-parameters are: $batch\_size = 32$, $lr = 7.3 \times 10^{-5}$, $weight\_decay = 0.001$, $dropout = 0.57$. The gradient is clipped at 3.21. The dimensions of the LSTM and corpus embedding are both 512.

The hyper-parameters for the model that includes backtranslated data are: $batch\_size = 32$, $lr = 1.2 \times 10^{-4}$, $weight\_decay = 1.52 \times 10^{-4}$, $dropout = 0.167$. The gradient is clipped at 1.84. The dimensions of the LSTM and corpus embedding are both 512.

C Results of Ostermann et al. (2017)

Ostermann et al. (2017)'s pipeline (1) identifies regular events from INSCRIPT and (2) classifies these regular events. The former was done with a J48 classifier, and the latter with a linear-CRF. To acquire its performance under our settings, we need to (a) take the output of their J48 classifier and (b) train and evaluate their second-stage model under our data split. As we were not able to find the implementation details of their J48 classifier, we only trained a second-stage model, using only the regular event, assuming perfect output from the J48 classifier. This model should perform strictly better than their complete pipeline. The numbers in Table 3 is a performance upper-bound for Ostermann et al. (2017), which suffices to validate our comparison. Furthermore, we trained our *Hierarchical* model under this simplified setting as well, and attained a $0.903/0.863$ micro/macro F1, which sees an even larger performance improvement.

References

James Bergstra and Yoshua Bengio. 2012. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. Allennlp: A deep semantic natural language processing platform.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Simon Ostermann, Michael Roth, Stefan Thater, and Manfred Pinkal. 2017. Aligning script events with narrative texts. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017)*, pages 128–134.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks

from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.