A Appendix : Technical Details

Our code is implemented with AllenNLP (Gardner et al., 2018) in Python 3. For BiDAF and QANet we use the default implementations from AllenNLP v0.8.2. [7] [8] For the QANet + Discourse-Aware Semantic Self-Attention we use the same hyper-parameters as QANet and we set the size of the label embeddings for the semantic information to $d_i = 16$.

A.1 Training Details

We train all models with Adam (Kingma and Ba, 2015) for up to 70 epochs with early stopping patience 20 and we halve the learning rate every 8 epochs if the Rouge-L on the *Dev* set does not improve. Depending on the model size and the used GPU, we train the models with effective batch size of up to 32. [9]

[7] BiDAF configuration https://github.com/allenai/allennlp/blob/v0.8.2/training_config/bidaf.jsonnet

[8] QANet configuration https://github.com/allenai/allennlp/blob/v0.8.2/training_config/qanet.jsonnet

[9] When training a model with the desired batch size does not fit on a single GPU, we use accumulated gradients as explained at https://medium.com/huggingface/training-larger-batches-practical-tips-on-1-gpu-multi-gpu-distributed-setups-ec88c3e51255.