A Model Hyper-parameters

Our hyper-parameters tuned on development set are: Adam optimizer with learning rate 0.05, maximum gradient norm 5.0, batch size is 32, embedding size 300, and hidden layer size of feed-forward network is 200 with dropout rate 0.1. The maximum vocabulary size is set to 30,000, but our dataset has a smaller vocabulary. The models are trained with either the original hypotheses or the sub-facts generated by ClausIE.

We test dropout ratio from 0.5, 0.75, 0.9 to 1.0, and encoder with glove averaging or LSTM. The maximum length of question and knowledge sentence is 25, and maximum length of supporting sentence in SciTail dataset is 40[5]. The hidden size of hybrid layer is 50, and hidden size of compositional layer is 50 and 2. The maximum number of knowledge per question is 100 and maximum number of sub-question per question is 5. The average number of sub-questions per question decomposed by (Del et al., 2013) is around 3.5. The learning rate is 0.05 and maximum gradient norm is 5.0 with Adam optimizer, and batch size is 32. We train our neural methods and NSnet network up to 25 epochs and choose the best model with validation and obtain accuracy on test set with the best model.

Based on the grid search over the hyperparameters, our best ENSEMBLE model uses *EmbOver* matcher on glove embeddings without tuple structure and probabilistic OR for hybrid decisions and averaging with 0.5 threshold for compositional decisions.

The best NSnet model uses *WordOver* matcher on glove encoding with tuple structure, no dropout ratio and sub-question training with neural models.

B Additional Experiments

For further analysis, we study effect of different matchers with(out) tuple structure, and different encoders (See Figure 3). The left figure shows test accuracies of symbolic (red) and NSnet (green) methods between three different lookup matchers (e.g., *EmbeddingAverage*, *EmbeddingOverlap*, *WordOverlap*) and whether tuple structure is considered (light) or not (dark). In most cases, *EmbeddingOverlap* that takes advantages from *EmbeddingAverage* and *WordOverlap*

[5]Science entailment dataset has long premise sentences

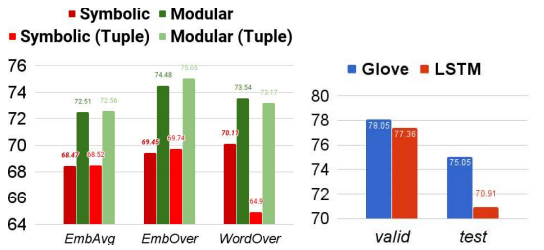

Figure 3: Comparison between different matchers with tuplization (left) and different encoders (right).

outperforms the others, and tuple structure helps for finding best matching knowledge tuple in our world knowledge base. The right figure shows accuracies between different encoders: averaging of glove word embeddings and LSTM with glove embedding initialization. LSTM is much worse in testing accuracy because of overfitting compared to glove embedding averaging.

C Observation on ENSEMBLE Model Design

For the ENSEMBLE network, we evaluated both OR and AND aggregation function and reported the best model. The use of AND is indeed intuitive. However, in addition to the empirical support for OR, the use of ClausIE to generate sub-facts makes probabilistic OR somewhat of a better fit, because of the following reason. ClausIE tries to generate every possible proposition in a sentence, erring on the side of higher recall at the cost of lower precision. This makes it unlikely for one to find good support for all generated sub-facts. This results in poor performance when using AND aggregation.