A The Hyper-parameters

We use the development dataset of ECB+ to tune the hyper-parameters of the proposed model HGCN in this work. The suggested values from the tuning process include: 2 layers for both sentence-level and document-level GCNs, 512 dimensions for the hidden vectors of the GCN models, 2048 dimensions for the hidden vectors of the feed-forward networks for the scoring functions (S_E, S_V), 16 for the minibatch size, $5e$-5 for the learning rate of the Adam optimizer, $\delta = 0.5$ for the predefined threshold of the agglomerative clustering algorithm (for both events and entities during the training and test phase), and $\alpha^{env} = 0.8$ and $\alpha^{ent} = 0.7$ for the trade-off parameters of the loss functions L_t^{env} and L_t^{ent}. Note that the tuning process suggests these hyper-parameter values for both EMLo or BERT embeddings in our model. Finally, we inherit the following resources from (Barhom et al., 2019) to ensure the compatibility: the same document clusters for the test phase that is computed from the Scikit-Learn tool (Pedregosa et al., 2011) (with 20 document clusters), the SwiRL tool for the SRL system (Surdeanu et al., 2007), and the Spacy tool for dependency parsing (Honnibal and Montani, 2017).

B Ablation Study for Entity Coreference

Table 4 reports the entity resolution performance on the ECB+ test set for the ablated/varied models of HGCN in the ablation study. The results in this table follow the trends of the model performance for event coreference resolution in Table 3, thus further demonstrating the benefits of the proposed components for HGCN.

Model	MUC	B^3	CEAF-e	CoNLL
HGCN (full)	**82.1**	**71.7**	**63.4**	**72.4**
HGCN-Sentence GCNs	80.3	71.0	62.9	71.4
HGCN-Pruned Tree	80.5	71.2	63.5	71.7
HGCN with One Sent GCN	78.2	70.6	63.2	70.7
HGCN-L_m^{reg}	80.3	70.9	62.0	71.0
HGCN-G^{doc}	80.7	71.0	63.8	71.8
HGCN-G^{doc}+TFIDF	80.2	70.2	63.3	71.2
HGCN-G^{doc}+MP	80.6	70.5	60.8	70.6

Table 4: The entity coreference resolution performance (F1) on the ECB+ test set.

C Error Analysis

We analyze the errors made by our HGCN model for CDECR to better understand its operation and provide some suggestions for future improvement. In particular, we sample 100 event mentions that are clustered incorrectly by the our model and manually categorize the error types. As such, following (Barhom et al., 2019), we consider a mention as being clustered incorrectly if its predicted cluster contains at least 70% of mentions that are not in their gold cluster. The following error types occur in our analysis (Figure 2 shows the distribution of such errors):

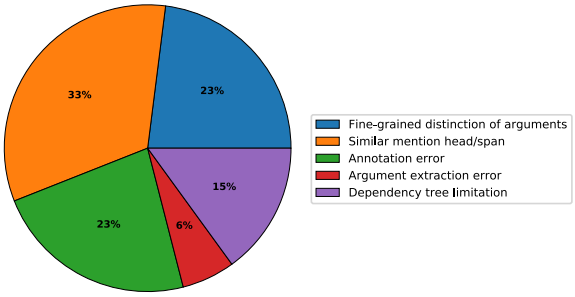

Figure 2: The error type distribution for our HGCN model on the ECB+ test set.

(i) Similar mention head/span (33%): This error of HGCN involves event mentions that are incorrectly grouped with other mentions that have similar mention heads/spans for the event triggers. This can be as simple as sharing the head lemmas or as complex as containing head words with similar meanings (e.g., "*deaths*" and "*killings*" in "*the* **deaths** *of a pregnant Arkansas woman*" and "*man guilty of* **killing** *pregnant girlfriend*"). This indicates that the current model has inappropriately reserved high weights for the event trigger head-related features in these cases and future work should consider other context information to better weight the information from the heads.

Note that given an event mention, the "Similar mention head/span" error type might co-exist with another error type in our analysis. As this is the loosest error type (i.e., only concerning mention heads/spans), we assign an event mention m to the other error type t if both "Similar mention head/span" and t apply for m.

(ii) Argument extraction error (6%): In this error, the SRL system incorrectly identifies event arguments for an event mention, causing HGCN to incorrectly disapprove the coreference of this event mention with another mention whose event arguments are correctly recognized. For instance, in the event mention (with "*hit*" as the trigger word) "*Wednesday 's shallow quake* **hit** *at 7 : 48 am (2248 GMT Tuesday) just off the coast , some 75 kilometres (50 miles) west of Manokwari.*", the SRL

system incorrectly predicts that "*Wednesday 's*" is an argument of the role `Arg0`. Here, the correct argument should involve "*shallow quake*". Our GCN system thus cannot make a correct coreference prediction for this event mention as "*Wednesday 's*" does not match any argument of the coreferring mentions.

(iii) Fine-grained distinction of arguments (23%): The failure of the model for this error is due to the close similarities of the arguments for event triggers, requiring fine-grained distinction between the arguments to correctly predict the coreference. For instance, the two following event mentions (with "*centered*" as the trigger words) are incorrectly assigned to the same cluster by HGCN:

A: "*The U.S. Geological Survey says the temblor at 9:27 a.m. was* **centered** *23 miles north of Santa Rosa.*"

B: "*The temblor at 2:09 a.m. was* **centered** *20 miles north of Santa Rosa, according to the U.S. Geological Survey.*"

As can be seen, the two event mentions have closely related arguments (e.g., the temblor, the north of Santa Rosa). To correctly reject the coreference in this case, the models should be able to distinguish the fine-grained difference between the arguments (i.e., "*9:27 a.m.*" vs "*2:09 a.m.*", and "*23 miles*" vs "*20 miles*").

(iv) Dependency tree limitation (15%): This error concerns the limitation of the shortest dependency paths between event triggers and arguments (with the four roles `Arg0`, `Arg1`, `Time`, `Location`) in identifying important context words for CDECR. As such, some important context words cannot be located using such shortest dependency paths, leading to errors for HGCN. For instance, in the two following event mentions with "*charged*" as the event triggers:

C: "*Jeffs is* **charged** *with two counts of sexual assault for raping two underage girls and fathering a child with one of the girls.*"

D: "*Five years ago, Warren Jeffs was* **charged** *with sex crimes resulting from the polygamous marriages he arranged for his followers in the Fundamentalist Church of Jesus Christ of Latter-Day Saints (FLDS).*"

HGCN incorrectly clusters these two event mentions into the same group partly due to the sole reliance on the shortest dependency paths between "*charged*" and the coreferring `Arg0` arguments/entity mentions "*Jeffs*" (in **C**) and "*Warren Jeffs*" (in **D**) to build the pruned trees for important context words. As such, the important context about the details of the charges in the two sentences have been missed (e..g, "*for raping two underage girls ...*" in **C** and "*resulting from the polygamous marriages he arranged for his followers ...*" in **D**), leading to the failure of the model in this case.

Note that "Fine-grained distinction of arguments" and "Dependency tree limitation" are exclusive as "Fine-grained distinction of arguments" only considers the context words in the noun phrase boundary for event arguments (to obtain the fine-grained distinction for only event mentions) while "Dependency tree limitation" concerns important context words beyond the event argument boundary. To address these two types of errors for CDECR, future work can explore better mechanisms to reveal important context words (both within and outside the noun phrases for event arguments) and emphasize on them for representation learning.

(v) Annotation error (23%): Similar to (Barhom et al., 2019), we also find annotation errors in our analysis where HGCN correctly predicts the coreference for some event mention pairs, but the annotation fails to record it.