A Multi-Attention based Neural Network with External Knowledge for Story Ending Predicting Task

Qian Li[1], Ziwei Li[1], Jin-Mao Wei[1], Yanhui Gu[2], Adam Jatowt[3], Zhenglu Yang[1]
CCCE, Nankai University,China[1]
School of CS and Technology, Nanjing Normal University, China[2]
Graduate School of Informatics, Kyoto University, Japan[3]
{liqian515,lzw_nku}@mail.nankai.edu.cn, weijm@nankai.edu.cn,
gu@jnu.edu.cn, adam@dl.kuis.kyoto-u.ac.jp,
yangzl@nankai.edu.cn

Abstract

Enabling a mechanism to understand a temporal story and predict its ending is an interesting issue that has attracted considerable attention, as in case of the ROC Story Cloze Task (SCT). In this paper, we develop a multi-attention based neural network (MANN) with well-designed optimizations, like Highway Network, and concatenated features with embedding representations into the hierarchical neural network model. Considering the particulars of the specific task, we thoughtfully extend MANN with external knowledge resources, exceeding state-of-the-art results obviously. Furthermore, we develop a thorough understanding of our model through a careful hand analysis on a subset of the stories. We identify what traits of MANN contribute to its outperformance and how external knowledge is obtained in such an ending prediction task.

1 Introduction

The prediction on story endings is an important and interesting application because it is involved with several essential issues, such as textual semantic understanding, logical reasoning and natural text generation. Most previous studies on the subject of common sense story understanding mainly focus on generating guesses for a missing event, such as matching explicit information in a given context (Chambers and Jurafsky, 2008), paying attention to specific types of common sense knowledge, like event schema (Chambers and Jurafsky, 2009), or concentrating on unsupervised learning (Chambers and Jurafsky, 2008). Although numerous studies have addressed the issue, training machines to be able to understand underlying narrative structures is still a challenging task. Previous research is limited at the shallow technique requirement of evaluation and noisy knowledge resources.

To facilitate the evaluation and benchmark the problem in the literature, the Story Cloze Task (SCT) has been introduced to predict what should be the "right" ending to a story (Mostafazadeh et al., 2016), which consists of daily events. The common strategies utilized to perform the task can be classified into two kinds of approaches: (1) traditional machine learning techniques with optimal feature engineering; and (2) deep learning-based models with effective strategies (Cai et al., 2017). The task is published with an unlabeled training set that consists of one-correct-ending stories, which is a notable impediment for further research. Some studies investigated fake ending generation, which obtained far more satisfying results. Most supervised learning approaches are trained from the finite evaluation set, ignoring the sheer volume of training corpus.

Understanding daily stories requires not only common sense experience sharing, but also a thorough understanding of text learned from common sense knowledge resources. We propose an effective multi-attention based neural network (MANN) and broaden our model with external knowledge. The model is superior in three aspects: (i) it adds features of sentences as embedding representations; (ii) it features a self-matched attention mechanism that functions through one sentence of the story, while interaction attention functions across the story plot and ending option to obtain word-level interaction; and (iii) it involves external knowledge to augment text coherence understanding.

Proceedings of the 27th International Conference on Computational Linguistics, pages 1754–1762
Santa Fe, New Mexico, USA, August 20-26, 2018.

The proposed model is comprehensively evaluated by comparing it with state-of-the-art approaches. Results show that the performance of the MANN strategy is superior to that of its competitors by up to 7% in terms of accuracy.

Furthermore, we conduct ablation experiments to illustrate the effect of each component in the MANN framework. Through elaborate evaluation, we demonstrate the superiorities and different characteristics of the components in the proposed model and the beneficial effect of the external knowledge utilized.

The contributions of this work are as follows:

- We build a MANN with deliberately devised structures, consisting of components that were not previously applied in this task, such as synthesis embeddings, multiple attentions and Highway, to characterize the semantic coherence of temporal stories from SCT.

- Unlike previous work on generating fake options or that trained only on the labeled set, we extend our model by regarding the unannotated corpus as proportions of external knowledge to enrich insufficient information to remedy the issue of limited resources.

- We conduct comprehensive analysis by manually labeling a subset of 300 stories to further study how our method performs in the story ending prediction task. We demonstrate that the proposed model outperforms the state-of-the-art methods by up to 7%. We will provide the full list of these annotated samples for further research.

2 Related Work

The issue of story ending prediction is related to several other research topics, such as reading comprehension and common sense learning, which will be briefly surveyed as follows.

Reading comprehension is the ability to read and understand text, and it has attracted much attention in natural language processing (NLP) to evaluate the level a machine can reach in understanding text. Two popular forms of evaluation tasks exist in this field: cloze-style query and text-span matching. Cloze-style query, such as SQuAD published by Stanford University, focuses on predicting existing text from the original corpus when given a relevant context. Text-span matching is different from selecting a possible word from the provided text to replenish the blank areas, such as CNN/DailyMail by Hermann and Hinton. Existing tasks are constructed with fragments, whereas examples from SCT are complete and independent stories that has short and meaningful sentence. SCT is also different in that it requires the prediction of development of a story, which is not provided in the given hypothesis. This novel task calls for stronger relation extraction and external inferential capability to identify the correct ending. Our model paid attention on through structure and proved to be effective during experiments.

Common sense learning is a challenging aspect in NLP. The limitation of other rich knowledge structures is that they mostly either focus on shallower representations, such as semantic roles like PropBank (Palmer et al., 2005), or pay attention to specific types of knowledge, i.e., unsupervised co-reference in the text (Chambers and Jurafsky, 2009) and event temporal relation (Modi and Titov, 2014). Learning from structural event knowledge is proposed to enrich this field, including narrative schema (Chambers and Jurafsky, 2009) and event frames (Sha et al., 2016). Unlike the above tasks, SCT (Mostafazadeh et al., 2016) provides large-scale supervised training stories of temporal and causal relations, ensuring a high-quality evaluation for common sense knowledge understanding of mechanisms.

However, the published ROCStories could not be used directly in supervised learning. Considering the use of the training set without negative endings, researchers proposed strategies to generate incorrect options. A conditional generative adversarial network has been proposed, achieving a moderate result with an accuracy of 60.9% (Wang et al., 2017). Roemmele (Roemmele et al., 2017) designed four generative models for fake options, namely, random, backward, nearest-ending and language model. The best result is produced from samples of all four types of endings (67.2%).

Other researchers have attempted to learn from the limit-scale validation set and augment the capability of the relation extractor. Schwartz (Schwartz et al., 2017) is the champion of the LSDSem 2017 Shared

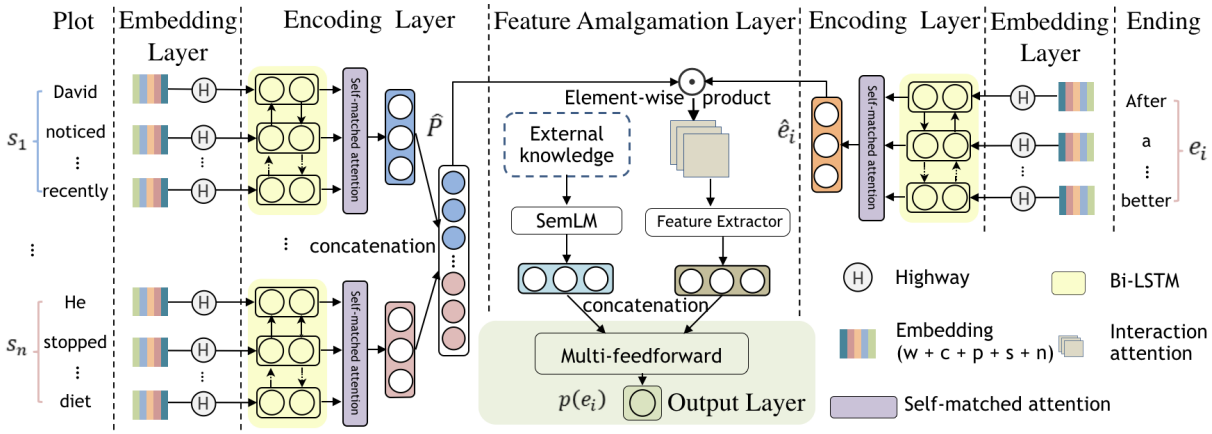


Figure 1: Architecture of our model.[1]

Task, which achieved a score of 75.2% by associating writing style features in endings and training a linear regression. HCM (Chaturvedi et al., 2017) trained a joint model with feature engineering to obtain representations of event sequence, sentiment, and topic from validation set and a hidden variable approach as a voter, thereby obtaining 77.6%. The previous NN-based models did not perform well. Cai (Cai et al., 2017) constructed a model with hierarchical long short-term memory network (LSTM) to encode plot and an ending2sentence attention, then concatenated the two representations through feed-forward network and outputting the final prediction, obtaining 74.7% accuracy. We pursue the same strategy to construct our principal model MANN and see opportunity to utilize external knowledge in the technique of combining semantic sequence information.

3 Proposed Model

This study aims to deduce the right ending given its previous context in the story. Formally, the story ending prediction task is defined as follows: given the story $\langle P, E\rangle$, where P=$\langle s_1, \ldots, s_n\rangle$ is a story plot, and the ending options E=$\langle e_1, \ldots, e_k\rangle$, the task is to select an appropriate ending e_i $(1 \leq i \leq k)$ from E. We address the task as a regression problem, mapping the right ending as 1 and the wrong ending as 0. We identify the highest-scored option as the answer. We first introduce the MANN model, followed with the extension of the model; this extension is constructed to learn from external knowledge. The whole model is shown in Fig.1.

3.1 Multi-attention Neural Network Model

The proposed MANN model consists of four layers. The embedding layer maps each word to a high-dimension vector representation. Then, the encoding layer encodes the representation of the context, and the feature amalgamation layer extracts features from the interaction between plot and the ending. Finally, the output layer provides the probability of the right ending.

Embedding Layer: We concatenate five representations: (i) word embedding; (ii) character feature; (iii) part-of-speech (POS) tagging; (iv) sentiment polarity of a word; and (v) negation.

Word embedding is conducted by converting a token to a high-dimensional vector space, and we obtain a fixed word embedding of each word by using pre-trained vocabulary. The character feature is obtained by mapping each word to a high-dimensional vector space to better handle out-of-vocab or rare words. We abstract character-level representations for tokens through a one-dimension convolutional neural network(CNN). The vectors embedded from characters have the same size as the input channel size of the CNN. By max-pooling the outputs of the CNN over the entire width, we obtain fixed-size vectors for words. By utilizing pre-processing tools, we tackle the last three aspects of a word as one-hot representations, while we collect the POS tagging feature with a natural language toolkit, sentiment polarity of a word with a look-up from pre-trained sentiment lexica, and negation with a corpus of

[1] w for word, c for character, p for POS, s for sentiment polarity, n for negation.

negation words (i.e., "not", "neither", "nor" and "n't"). The concatenation of the five representations of a word is then passed through a two-layer Highway Network (Srivastava et al., 2015) to fuse the information of features, which is processed as follows:

$$trans = ReLU(w_t x + b_t) \tag{1}$$

$$gate = \sigma(w_g x + b_g) \tag{2}$$

$$H(x) = gate * trans + x(1 - gate) \tag{3}$$

where $w_g, w_t \in \mathbb{R}^{D\times D}$ and $b_t, b_g \in \mathbb{R}^D$, D is the dimension of input.

Encoding Layer: The encoding layer encodes the sequence and semantic abstraction of a single sentence and then compromises them, which are from an identical plot to obtain a fusion premise.

Sentence encoding constructs an LSTM in both directions on top of the embeddings provided by the previous layer and concatenate the outputs of forward and backward LSTMs, to learn high-level abstractions from time-sequence features of the context. We obtain $s_{lj,(j=1,\ldots,n)} = [\overrightarrow{LSTM}; \overleftarrow{LSTM}]$ where ; represents the concatenation between two directional LSTM thus $s_{lj} \in \mathbb{R}^{T\times 2d}$ denotes each sentence in plot and $e_{li} \in \mathbb{R}^{G\times 2d}$ denotes the ending.

The new representations are passed into self-matched attention to model the temporal interactions between words. Taking vector V as example, self-matched attention $\widetilde{V} = att(V)$ is defined as follows:

$$M_{ij} = W^{\mathrm{T}}[V_i; V_j; V_i \circ V_j] \tag{4}$$

$$a_i = softmax(M_i) \tag{5}$$

$$\widetilde{V_i} = \sum_j a_{ij} V_j \tag{6}$$

where $W^T \in \mathbb{R}^{6d}$ is a weight matrix and $\circ$ is element-wise multiplication. The higher-level semantics can directly tackled from encoded sequences through the attention mechanism.

Plot encoding compounds the interacted representation of sentences from the same story context. We studied several commonly used implementations for sentence combination, such as LSTM, summation, weighted summation, and concatenation operations. We find simple concatenation useful. The encodings of the story plot is concatenated as $\hat{P} = [att(s_{l1}); \cdots ; att(s_{ln})]$, $\hat{P} \in \mathbb{R}^{nT\times 2d}$, while the ending is simply represented as $\hat{e} = att(e_{li})$, $\hat{e} \in \mathbb{R}^{G\times 2d}$.

Feature Amalgamation Layer: We focus on characterizing diverse interaction information among representations of the plot and the ending, thus extracting feature from them. This layer is inspired by the IIN model (Gong et al., 2018). We combine the plot vectors and ending representations to create a word-by-word interaction attention tensor, in which each channel represents the interaction of the word in one dimension. We tried processing, such as $F_{ij} = \hat{P}_i \circ \hat{e}_j$, $F_{ij} = \hat{P}_i + \hat{e}_j$ and $F_{ij} = |\hat{P}_i - \hat{e}_j|$ and found that the most effective one is $F_{ij} = \hat{P}_i \circ \hat{e}_j$, where we define $i \in [1, \ldots, nT]$, $j \in [1, \ldots, G]$, $\circ$ is element-wise multiplication and $F \in \mathbb{R}^{nT\times G\times 2d}$.

Then, we adapt a feature extractor on F to extract semantic features from word-by-word interaction.Unlike extractors (i.e., VGG and ResNets (He et al., 2016)), DenseNet (Huang et al., 2017) strengthens feature propagation and reduces information disappearance through time because of the structure of pre-activation.

Output Layer: A multi-feedforward neural layer is used. We apply three $tanh$ layers to calculate a score for prediction support.

3.2 Extension of MANN

Labeling large-scale examples requires considerable expertise and manpower. With the pattern of supervised learning, MANN can merely use limited annotated examples to train finite information, thereby preventing us from obtaining more powerful statistical model. To eliminate the restriction of labeled data scarcity and ensure the robustness of our model, we introduce semantic sequence information extracted from external knowledge onto MANN, thus building an extended model, i.e., sequence based MANN (SeqMANN).

External knowledge is mostly used to enrich word implication, thereby ensuring that semantic information can be extracted. To address the combination of external knowledge and the neural network model, Chen (Chen et al., 2017) added extra relation informations of word pair into the encoder. Other researchers studied embedding representations (Bordes et al., 2013) to learn complex reasoning capacities. Considering specific ending-prediction task, we aim to further derive coherence among sentences from stories. Directly working on the large-scale ROCStories is not easy due to its deficiency of negative samples. We use SemLM (Peng and Roth, 2016) to model the distribution over a meaningful sequence chain, with ROCStories regarded as portion of our extra resources while the other are from news data.

SemLM is a language model that first uses FrameNet to split sentences by semantic sequence, and then represents these pieces of sequence with semantic frames and discourse markers from an extended vocabulary. The abstraction of a sentence is $[f_1, dis_1, f_2, \cdots, o]$ where f_i denotes semantic frames, dis_i denotes discourse markers and o denotes period symbol. The SemLM is trained with a log-bilinear model (Mnih and Hinton, 2007) on ROC corpus and news data, and obtains the probability of two words appearing simultaneously in a sentence. The log-bilinear model computes the sequence probability of the next word w_i given the previous words (context), which is defined as follows:

$$p(w_i|c(w_i)) = \frac{exp(v(w_i)^{\mathrm{T}} u(c(w_i)) + b(w_i))}{\sum_{w \in V} exp(v(w)^{\mathrm{T}} u(c(w_i)) + b(w))} \tag{7}$$

We define $v(w)$ as the target vector, $v'(w)$ as the context vector, and $b(w)$ as a bias of a token. Here, V is the vocabulary, $u(c(w_i)) = \sum_{c_t \in c(w_i)} q_t \circ v'(c_t)$, $\circ$ is element-wise multiplication and q_i is a model parameter that depends on the position of a token in the context. The final sequence probability is $\prod_{i=1}^{k} p(w_i|c(w_i))$.

The trained language model is then used to calculate the conditional probability of semantic frames from each option when given the same hypothesis, inspired by HCM (Chaturvedi et al., 2017). For each ending e_i with frames represented as f_{e_i} and $\langle f_1, f_2, \cdots, f_T \rangle, T \geq n$ indicating semantic frames evoked in the story plot, the following features are captured considering the sequence of frames in corresponding story plot: $P(f_{e_i}|f_T), P(f_{e_i}|f_T, f_{T-1}), \ldots, P(f_{e_i}|f_T, f_{T-1}, \cdots, f_1)$.

Finally we learn the semantic sequence feature of plot-ending pairs which represents the interaction information between the plot and the ending of stories, extending the study on external knowledge. We fuse it with features extracted from DenseNet in the feature amalgamation layer through concatenation.

4 Experiments

4.1 Data

As described in (Mostafazadeh et al., 2016), SCT was constructed based on ROCStories. The ROC corpus consists of 100,000 five-sentence cases, each of which was written as a logically meaningful story. After eliminating original endings, writers develop both a "right" ending and a "wrong" ending for the context of examples which are randomly chosen from the corpus. The published SCT is constructed with ROCStories as a large training set, an evaluation set, and a test set, which have the same structure and a size of 1,871.

We evaluate our model by using the benchmark SCT (Mostafazadeh et al., 2016). Notably, the training set contains four-sentence articles with one correct ending, while the evaluation set consists of four-sentence stories with two ending options.

4.2 Training details

To train our neural algorithm, we apply word embeddings of a look-up from 100-d GloVe pre-trained on Wikipedia and Gigaword (Pennington et al., 2014). We set $hiddensize = 100$ for LSTM. An Adam optimizer with a mini-batch size of 120 and an initial learning rate of 0.01 is applied. In the feature amalgamation layer, DenseNet consists of three pairs of dense blocks with a following transition block. The number of layers in a dense block is set as 10, and a ReLU activation function is applied for the whole convolutions. In the output layer, we use three full-connected layers with ReLU activation function. We use mean squared error as the loss function. We decide the model based on the average accuracy of the held-out folds through 5-fold cross validation.

Table 1: Performance comparison.

	Acc.
Machine Learning Algorithm	
Acoli (Schenk and Chiarcos, 2017)	70.0%
Schwartz (Schwartz et al., 2017)	75.2%
HCM (Chaturvedi et al., 2017)	77.6%
Neural Network Algorithm	
DSSM (Mostafazadeh et al., 2016)	58.5%
CGAN (Wang et al., 2017)	60.9%
Lin (Lin et al., 2017)	67.0%
LSTM (Mihaylov and Frank, 2017)	72.8%
Cai (Cai et al., 2017)	74.7%
MANN	78.3%
SeqMANN	84.7%
Human Performance	100%

Table 2: Ablation study.

	MANN	SeqMANN
Embedding Ablation		
-character feature	76.5%	84.3%
-sentiment polarity	75.7%	84.3%
-POS	77.5%	84.5%
-negation	76.3%	84.4%
-word embedding	73.4%	82.9%
Component Ablation		
-Highway	76.3%	83.0%
-biLSTM	74.3%	82.3%
-self_matched attention	75.1%	82.4%
-interaction attention	74.6%	80.0%
-MANN	—	71.3%
Full	78.3%	84.7%

4.3 Results on Our Model

We compare our model with some distinctive methods and approaches that rank high in LSDSem 2017 Shared Task (Mostafazadeh et al., 2017) in Tabel 1. Cai (Cai et al., 2017) is state-of-the-art model in NN-methods while HCM (Chaturvedi et al., 2017) is state-of-the-art model in all published methods. Under the condition in which external resources are not used, MANN outperforms Cai by 3% and even performs better than the highest-level method. Benefiting from external knowledge, SeqMANN achieves a 6.4% improvement over MANN. Nevertheless, the superiority of SeqMANN does not rely on only the external resources utilized, but the correlated effect between MANN and the external resources, as we can see that after removing MANN from SeqMANN (as shown in Table 2 and will described in Section 4.4), the remaining external resources based model only obtained 71.3%[1].

4.4 Ablation Study

We conduct an ablation study on the proposed model to evaluate the effectiveness of each feature and component involved. Results are shown in Table 2.

Embedding Ablation: All embedding features contribute to MANN. However, the influence is not as crucial in SeqMANN. We conjecture that contextual information in the embedding process partly overlaps with the features in external knowledge.

Component Ablation: We respectively remove each component to study their contribution. For interaction attention ablation, we replace it with flattening between plot and ending representations, followed by removing DenseNet and replacing a three-layer $tanh$ operation as feature extractor. For MANN ablation, we only retain the extraction of external knowledge and the output layer (we model SemanticSequence by the same way as MANN ablation processing in Section 5).

5 Data Analysis

To further analyze the model, we seek to determine the abilities required to predict the right ending, and which aspects of questions among the specific dataset are solved by our model. We sample 300 examples from the validation set randomly and annotate them manually from two cognitive chunks, namely, by labeling samples with the difficulty degree in human-understandable rationale for prediction and by tagging samples according to the linguistic phenomena they contain.

[1]Code and the annotated samples are available at https://github.com/StoryDevelopment/SCT.

Table 3: Some examples from each human-understandable difficulty labeling category.

Category	Story plot	Ending options
Relevant Word	Kyle invited everyone he works with bowling one night. Most people could not go but Matt and John showed up. Matt had never been bowling before so they had to show him how to. After a few games, Matt picked up how to play better.	e_1: Now Matt and Kyle can go bowling more then. e_2: Kyle took the children shopping for a gift for their mother.
Compatible Sentence	It was the last day of our vacation.We were eating lunch on the patio of the hotel. We laughed and smiled because it was a great vacation.Then we packed our bags and drove to the airport.	e_1: We want to revisit someday. e_2: We all vowed to never go back again.
Plot-level Paraphrasing	Rory was allergic to gluten and strawberries.One day she sat down to eat lunch at school.She opened her lunch box, and stared at a sandwich with strawberries.Her new step mom had packed her lunch for the first time.	e_1: Rory had to buy a school lunch that day. e_2: Rory ate the sandwich.
Ambiguous Inference	Tina always wore a red bikini when she went to the beach. She was known for it and everyone expected to see her in one. One day she met her friends in a blue bikini and surprised them. They could not understand why she would wear something different.	e_1: They liked the new bikini though. e_2: Tina's friends knew how unpredictable she was.

5.1 Human-understandable Difficulty Labeling

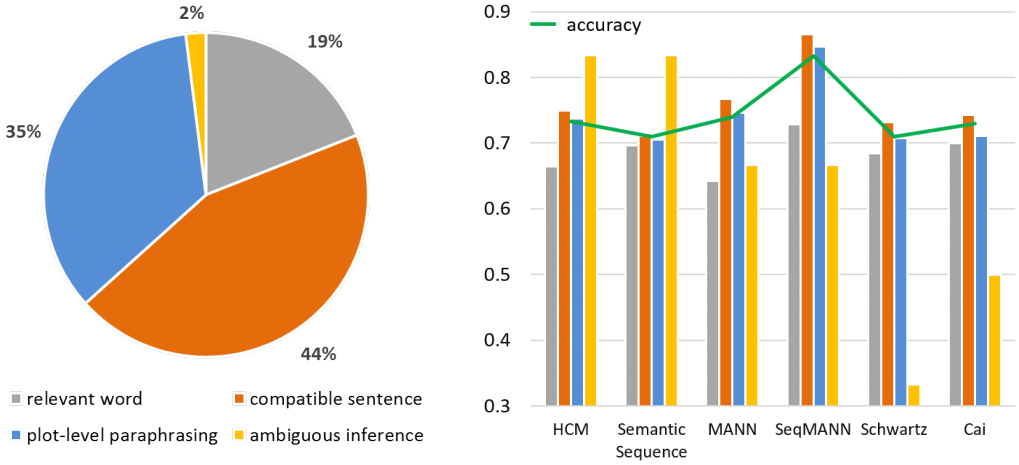
We classify the 300 samples into the following categories (as shown in Table 3, where the right ending is e_1 and the wrong ending is e_2):

Relevant Word. The verbs and noun phases in e_1 are more relevant to the plot than those in e_2. This category includes examples that are thought to be correctly predicted at the word level.

Compatible Sentence. We could derive the correct answer by understanding a single sentence in the hypothesis, which means that the answer is similar to some sentence in an earlier context.

Plot-level Paraphrasing. It requires a full understanding of multiple sentences to infer the answer.

Ambiguous Inference. It includes examples in which we think both endings are logically reasonable for the story, while the correct one is more answerable. This category consists of poorly designed cases, which we consider miscarriage examples in this task.



(a) The proportion of each category.

(b) Results in each category.

Figure 2: Results of human understandable difficulty labeling.

Fig.2a shows the proportion of each category. The second and third cases have the highest proportion, followed by the first and the least "Ambiguous Inference" cases. We observe that deeper information processing, which means sentence understanding (for "Compatible Sentence" case) and plot-level understanding (for "Plot-level Paraphrases" case), is crucial to good performance in this task. The low proportion of "Ambiguous Inference" shows the satisfactory quality of this task.

To further analyze the depth of prediction from our models compared with other methods, we reproduce the other methods and test the 300 annotated examples based on the above categorization. Results are presented in Fig.2b. We conclude the following: (i) The comparison between MANN and semantic sequence shows that semantic sequence is superior in shallow natural language comprehension, such as word-level understanding, while MANN performs better in deep contextual comprehension, such as sentence and plot level understanding; (ii) Successfully, SeqMANN combines the two advantages to out-

Table 4: Linguistic phenomena tagging results.

Phenomena Tag	Label Frequency	Cai	Schwartz	HCM	SemanticSequence	MANN	SeqMANN
CONDITIONAL	2.0	83.3	83.3	83.3	66.7	66.7	**100.0**
NEGATION	42.7	70.3	67.9	69.5	64.1	70.3	76.6
SENTENCE LENGTH	42.0	76.1	76.5	75.3	70.3	72.2	81.7
QUANTITY/TIME REASONING	21.0	73.0	60.3	66.7	65.1	74.6	82.5
COREF	39.3	83.1	75.4	78.0	75.4	77.1	**88.1**
QUANTIFIER	31.3	73.4	69.1	67.0	61.7	74.6	82.5
MODAL	27.0	75.3	76.5	75.3	70.3	69.1	80.2
BELIEF VERBS	5.0	53.3	53.3	60.0	60.0	73.3	73.3
CONVERSATIONAL PIVOTS	31.3	76.6	69.1	74.5	75.5	70.0	**84.0**
ANTO	24.7	73.5	66.3	74.7	67.5	68.7	83.1
EMOTIONAL COMMONALITY	68.7	73.7	72.3	74.3	71.3	78.2	**86.9**
ENDING ONLY	4.0	58.3	76.5	60.0	60.0	91.6	**83.3**
ACCURACY	—	72.3	71.3	73.3	71.0	74.0	83.3

perform other methods on this task. (iii) The model is a stable one with external knowledge resources, thus still maintaining high accuracy under decreasing training data.

5.2 Linguistic Phenomena Tagging

To obtain an idea of the linguistic phenomena in SCT and to conduct a detailed analysis of the semantical performance of these models, imitating MultiNLI (Williams et al., 2018), we design a set of annotation tags to label the subset as follows:

CONDITIONAL: Whether the example contains the word "if".
NEGATION: Whether the example contains negation words, e.g., not, none, neither.
SENTENCE LENGTH: Whether the right ending is longer than the other endings.
QUANTITY/TIME REASONING: Whether understanding the ending options contains quantity or time reasoning that needs to be explained from the plot.
COREF: Whether ending options contain referring expressions.
QUANTIFIER: Whether the example contains quantifier words, e.g., more, most, enough.
MODAL: Whether the example contains modal verbs.
BELIEF VERBS: Whether the example contains belief verbs such as think, believe and doubt.
CONVERSATIONAL PIVOTS: Whether the example contains discourse cohesion, e.g., but, yet, however, though, while.
ANTO: Whether the example contains an antonym pair.
EMOTIONAL COMMONALITY: Whether the sentiment throughout the plot is consistent with that through the right ending.
ENDING ONLY: Whether the ending options contradict themselves.

Results are shown in Table 4, and we observe the following: (i) The poor performances on SENTENCE LENGTH indicates that such feature, which is regarded as valuable bias (Cai et al., 2017; Schwartz et al., 2017), does not have an influential contribution to our model. (ii) Examples with ENDING ONLY tag shows that our attention components recognize not only paraphrases in relation among sentences but also the intra-semantic implications of each sentence. (iii) The model outperforms in terms of EMOTIONAL COMMONALITY, which shows effects of adding sentiment polarity in embedding representation.

6 Conclusion

In this paper, we proposed a MANN model with external knowledge. The results of this model outperformed state-of-the-art results by 7%. We carefully examined a subset of the corpora from SCT to analyze the performance of our models. The competitive model benefits not only from our thoughtfully designed structure but also from the combination of semantic relations learned from external resources.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant No.U1636116, 11431006, 61772288, the Research Fund for International Young Scientists under Grant No. 61650110510 and 61750110530, and the Ministry of education of Humanities and Social Science project under grant 16YJC790123.

References

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *NIPS*, pages 2787–2795.

Zheng Cai, Lifu Tu, and Kevin Gimpel. 2017. Pay attention to the ending: Strong neural baselines for the roc story cloze task. In *ACL*, pages 616–622.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *ACL*, pages 789–797.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *ACL and AFNLP*, pages 602–610.

Snigdha Chaturvedi, Haoruo Peng, and Dan Roth. 2017. Story comprehension for predicting what happens next. In *EMNLP*, pages 1603–1614.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, and Diana Inkpen. 2017. Natural language inference with external knowledge. *arXiv preprint arXiv:1711.04289*.

Yichen Gong, Heng Luo, and Jian Zhang. 2018. Natural language inference over interaction space. In *ICLR*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*, pages 770–778.

Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. 2017. Densely connected convolutional networks. In *CVPR*, pages 4700–4708.

Hongyu Lin, Le Sun, and Xianpei Han. 2017. Reasoning with heterogeneous knowledge for commonsense machine comprehension. In *EMNLP*, pages 2032–2043.

Todor Mihaylov and Anette Frank. 2017. Story cloze ending selection baselines and data examination. In *LSDSem*, pages 87–92.

Andriy Mnih and Geoffrey Hinton. 2007. Three new graphical models for statistical language modelling. In *ICML*, pages 641–648.

Ashutosh Modi and Ivan Titov. 2014. Inducing neural models of script knowledge. In *CoNLL*, pages 49–57.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and evaluation framework for deeper understanding of commonsense stories. In *NAACL*, pages 839–849.

Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. 2017. Lsdsem 2017 shared task: The story cloze test. In *LSDSem*, pages 46–51.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

Haoruo Peng and Dan Roth. 2016. Two discourse driven language models for semantics. In *ACL*, pages 290–300.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543.

Melissa Roemmele, Sosuke Kobayashi, Naoya Inoue, and Andrew M Gordon. 2017. An rnn-based binary classifier for the story cloze test. In *LSDSem*, pages 74–80.

Niko Schenk and Christian Chiarcos. 2017. Resource-lean modeling of coherence in commonsense stories. In *LSDSem*, pages 68–73.

Roy Schwartz, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, and Noah A Smith. 2017. The effect of different writing tasks on linguistic style: A case study of the roc story cloze task. In *CoNLL*, pages 15–25.

Lei Sha, Sujian Li, Baobao Chang, and Zhifang Sui. 2016. Joint learning templates and slots for event schema induction. In *NAACL-HLT*, pages 428–434.

Rupesh K Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Training very deep networks. In *NIPS*, pages 2377–2385.

Bingning Wang, Kang Liu, and Jun Zhao. 2017. Conditional generative adversarial networks for commonsense machine comprehension. In *IJCAI*, pages 4123–4129.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL*.