Find a Reasonable Ending for Stories:
Does Logic Relation Help the Story Cloze Test?

Mingyue Shang,[1,◇] **Zhenxin Fu,**[1] **Hongzhi Yin,**[2] **Bo Tang,**[4] **Dongyan Zhao,**[1,3] **Rui Yan**[1,3*]

[1]Institute of Computer Science and Technology, Peking University, China, [2]The University of Queensland, Australia
[3]Center for Data Science, Peking University, China, [4]Southern University of Science and Technology, China
◇Email: shangmy@pku.edu.cn

Abstract

Natural language understanding is a challenging problem that covers a wide range of tasks. While previous methods generally train each task separately, we consider combining the cross-task features to enhance the task performance. In this paper, we incorporate the logic information with the help of the Natural Language Inference (NLI) task to the Story Cloze Test (SCT). Previous work on SCT considered various semantic information, such as sentiment and topic, but lack the logic information between sentences which is an essential element of stories. Thus we propose to extract the logic information during the course of the story to improve the understanding of the whole story. The logic information is modeled with the help of the NLI task. Experimental results prove the strength of the logic information.

Introduction

Natural language understanding is an important field of Natural Language Processing which contains various tasks such as text classification, natural language inference (NLI) and story comprehension. These tasks differ in many aspects including the input formation and task goals. For example, the text classification task takes one sequence (a sentence or a document) as input to predict a label, while the NLI task takes two sequences as inputs to infer a relationship. Previous methods typically adopt supervised training to train a task-specific model. As the goal of tasks varies, the features that each model could extract also differs. While these features are treated separately in previous methods, it is intuitive that combining features across tasks may help to enhance the information representations. In this paper, we aim to explore the effectiveness of incorporating the features from one specific task to another task.

Particularly, we investigate the feasibility of incorporating features learned in the NLI task into Story Cloze Test (SCT) (Mostafazadeh et al. 2016) for the following reasons: datasets of these two tasks are in similar but still subtly different domains; the input and output form of the tasks are completely different which means distinct features are learned during the supervised training process. The NLI task

---

predicts a relation from {*Entail, Neutral, conflict*} of two sentences. The SCT is designed to choose a reasonable ending from two candidates, given four sentences as context (a.k.a plot). Compared with SCT, labels in the NLI task indicates explicit logic relations. In this paper, we explore to enhance the logic information representations with the help of NLI task to infer the relationship between the story context and the ending, as logic is an essential element of a story.

To this end, we propose a fully neural network based method that contains three components: a *Content Unit* (CU) to encode the plot and the ending to get a content vector of the whole story; a *Logic Unit* (LU)–the core unit to introduce logic features–to extract the logic relationship between each sentence in the plot and the ending, and then track the logic flow of the whole story; a *Score Unit* (SU) to generate a score to indicate the reasonableness of the ending for the plot. Drawing upon modeling both content and logic into SCT, our method aims to provide new insights into story understanding task.

Architecture

Figure 1 gives an overview of the proposed model. Concretely, a plot $p = \{w_1, \ldots, w_4\}$ and a candidate ending $e = w_5$ are fed to both CU and LU, where w_i means the i-th sentence. After that, the SU takes the output of CU and LU as input and produces a score. A higher score means the ending is more reasonable to the plot, and vise versa.

Content Unit. Inspired by Cai, Tu, and Gimpel (2017), we implement the hierarchical recurrent neural network (RNN) to encode the 4-sentence plot and the ending into a content vector. The hierarchical RNN contains a word level RNN (referred as W-Enc) and a sentence level RNN (referred as S-Enc), which are bi-directional GRU for both levels.

The W-Enc first converts each sentence w_i into hidden states $h_{w,i} = \{h_{w,i}^1, \ldots, h_{w,i}^n\}$, and then computes the attention of the ending over each sentence through hidden states, denoted as $\beta_i = \{\beta_i^1, \ldots, \beta_i^n\}$. The sentence vector s_i is calculated as the attention-weighted sum of the hidden states. Then the S-Enc encodes the plot $P = \{s_1, \ldots, s_4\}$ and the ending $E = \{s_5\}$, respectively. The final content vector V_c is defined as $V_c = [u_p; u_e]$, where u_p and u_e are the hidden state of the plot P and the ending E at the final step, and $[;]$ means concatenation.

Logic Unit. Logic Unit consists of a logic extractor to capture the logic relation of two sentences and a logic tracker to
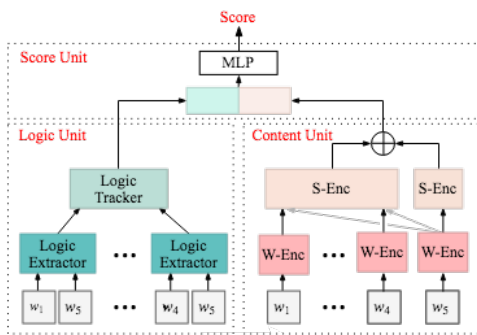
Figure 1: The overview of the proposed model which contains three main components.

track the sequential logic information of the story.

Logic extractor. For the purpose of enhancing logic information purpose, we pre-train an NLI model on NLI dataset. The NLI model follows the structure of ESIM (Chen et al. 2017). Readers can refer to Chen et al. (2017) for more details. During the pre-training process, the ESIM takes two sentences as inputs and encodes them to vector representation, which is then fed to a multi-layer perceptron (MLP) with softmax operation to predict a label. The parameters are updated by cross-entropy loss. In this work, we discard the final layer of the MLP in the ESIM and employ the remainder as the extractor. When trained on the SCT, each sentence in the plot is coupled with the ending to form a sentence pair p_i. Given the pair as input, the extractor gives an intermediate vector $u_{l,i}$ which represents the logic information.[1]

Logic Tracker. Build upon the extractor, the logic tracker aims to model the logic flows of the whole story. Four sentence pairs are fed to the extractor respectively, and the outputs of the extractor form a sequence $L = \{u_{l,1}, \ldots, u_{l,4}\}$. The tracker is implemented by a BiGRU to encode L into hidden states $h_t = \{h_{t,1}, \ldots, h_{t,4}\}$. Then the sequential hidden states are concatenated as the final output of the logic tracker $V_l = [h_{t,1}; \ldots; h_{t,4}]$.

Score Unit. The final prediction is based on the combination of content information V_c and the logic information V_l. They are concatenated and fed into a score function to predict a score S which measures how reasonable the ending is. The score function S is implemented by a MLP with $\tanh$ activation function.

As each plot is accompanied with a right ending and a wrong ending, we apply hinge loss to train the model. Concretely, S^+ denotes score of the story with the right ending and S^- denotes the score with the wrong ending. The objective is to minimize L, where $L = max(0, \ \Delta - S^+ + S^-)$, Δ is the margin and is set to 1 in this paper.

Experiments

Dataset. We conduct the story cloze experiments on the ROCStories corpus (Mostafazadeh et al. 2016) which has

---

[1]As the model is trained by the labels that requires the model to extract the logic relation, the output before the softmax layer in the MLP is supposed to contain the logic information as to correctly predict the label.

Model	ACC.
HEPA	74.7%
LS-Skip	76.5%
HCM	77.6%
FTLM	**86.5**%
Our Model	79.1%

Table 1: Test-set results.

Model	ACC.
CU	71.7%
LU	75.6%
All	79.1%

Table 2: Ablation study.

98,161 stories in the training set and 1,871 4-sentence cases in the validation set and test set respectively. Each 4-sentence case has a right ending and a wrong ending. The ESIM mentioned in the logic extractor is trained on an NLI dataset. Further details are provided in the supplement[2].

Experimental Results. Table 1 shows the results of our approaches and other models . Among the reported baselines, HEPA (Cai, Tu, and Gimpel 2017) and LS-Skip (Srinivasan, Arora, and Riedl 2018) are fully neural network based, but their performances are slightly weaker then others. The Hidden Coherence Model (HCM) (Chaturvedi, Peng, and Roth 2017) relies on the feature engineering. Though our performance is weaker than the current state-of-the art result achieved by the FTLM (Radford et al. 2018), compared with FTLM which requires large-scale pretrain, our model involves less data and focuses on investigating the use of logic information. The ablation results further verify the effectiveness of the cross-task information in Table 2.

Conclusion and Future Work

In this paper, we aim to investigate the feasibility of incorporating cross-task information. We conduct experiments on SCT with the enhancement from the NLI task. In the future, we plan to explore the possibility of inter-task combination from other aspects.

References

Cai, Z.; Tu, L.; and Gimpel, K. 2017. Pay attention to the ending: Strong neural baselines for the roc story cloze task. In *ACL (Short Papers)*, volume 2, 616–622.

Chaturvedi, S.; Peng, H.; and Roth, D. 2017. Story comprehension for predicting what happens next. In *EMNLP*.

Chen, Q.; Zhu, X.; Ling, Z.-H.; Wei, S.; Jiang, H.; and Inkpen, D. 2017. Enhanced lstm for natural language inference. In *ACL (Long Papers)*, volume 1, 1657–1668.

Mostafazadeh, N.; Chambers, N.; He, X.; Parikh, D.; Batra, D.; Vanderwende, L.; Kohli, P.; and Allen, J. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of NAACL-HLT*, 839–849.

Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. Improving language understanding by generative pre-training.

Srinivasan, S.; Arora, R.; and Riedl, M. 2018. A simple and effective approach to the story cloze test. In *NAACL-HLT*.

---

[2]https://shangmy.github.io/publication/story-cloze