Joint Abductive Generation and Discrimination via Cycling Reasoner

Xin Zheng
Chinese Information Processing Laboratory
Institute of Software, Chinese Academy of Sciences, Beijing, China
University of Chinese Academy of Sciences, Beijing, China
zhengxin2020,@iscas.ac.cn

Abstract

Abductive reasoning, one of the core human cognition abilities, enjoys rapid progress separately in its two task forms: Abductive Natural Language Inference (αNLI) and Abductive Natural Language Generation (αNLG). However, traditional teacher-forcing learning strategies may not address the diverse issue and subtle semantic differences of abductive reasoning very well, undermining the performance on both tasks.

In this paper, we propose Cycling Abductive Reasoner (CAR), which jointly exploits αNLI and αNLG, to resolve the challenges. We leverage the duality between the classification and generation tasks, first filter αNLG generated hypothesis by αNLI model for better coherence, then for more diversity, use the enhanced generation process to produce novel instances for αNLI discriminator, and finally the augmented αNLI model is more robust at discriminating the diverse hypotheses. Experiments show that our approach achieves new State-of-the-Art performance on both αNLI and αNLG tasks when submitted to the AI2 Leaderboard[1], with a noticeable improvement on αNLG human evaluation score, demonstrating the effectiveness of our approach.

1 Introduction

Abductive Reasoning (Peirce, 1974) is to find the plausible explanations given limited observations, which is a fundamental intelligence ability. For example, given a premise "Jenny cleaned her house and went to work, leaving the window just a crack open." and a conclusion "When Jenny returned home, she saw her house was a mess.", abductive reasoning requires to generate and evaluate potential hypothesis like "A thief broke into the house by pulling open the window." happens between them.

Currently, two kinds of basic tasks are proposed to evaluate the machine's ability to conduct abductive reasoning. One is Abductive Natural Language Generation (αNLG), which is to generate a plausible hypothesis given two incomplete observations. Another is Abductive Natural Language Inference (αNLI), which aims to discriminate whether a hypothesis is reasonable in the same setting. For example, given a past observation "Rob's school was holding a Halloween costume contest" and a future observation "He was proud of himself, and decided to do even better next year", αNLG requires the machine to generate potential plausible hypotheses like "He made a costume and it ended up winning second place", while αNLI requires the machine to discriminate whether a given hypothesis is reasonable or not.

Previous works commonly address these two tasks individually, learning αNLI or αNLG models by discriminating or generating golden hypotheses in the existing corpus using supervised learning models. Unfortunately, such independent teacher-forcing learning strategies may contradict the nature of abductive reasoning. Due to the incomplete information and uncertainty of real world, multiple significantly divergent hypotheses might be correct for the same past and future observations. Moreover, two hypotheses with minor surface form differences might yield significant differences in terms of plausibility. For example, "won the second place" and "won the first place" shares similar surface utterance and semantics, but they will lead to very different plausibility when put in the contexts of "A Halloween costume contest" and "Decided to do even better next year". Consequently, such teacher-forcing learning strategies undermine both αNLI and αNLG performance: αNLI model might not robustly discriminate when facing unseen reasonable hypothesis, and αNLG might be challenging to generate diversified hypothesis and distinguish similar utterances when multiple possibilities exist.

In this paper, we proposed Cycling Abductive

[1] https://leaderboard.allenai.org/

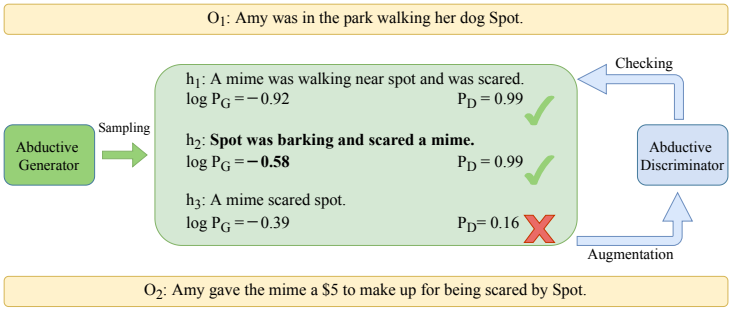

Figure 1: Illustration of Abductive Reasoning Cycle.For the data augmentation of αNLI, we generate novel hypotheses by the generator and reassured by αNLI; at test time of αNLG, we use augmented αNLI discriminator to throw out implausible hypotheses like h_3;

Reasoner (CAR), a novel method that jointly exploits αNLI and αNLG to resolve the challenges mentioned above. The main idea behind CAR is to explore the duality of αNLI and αNLG, i.e., these two tasks share the same underlying abductive reasoning abilities and, therefore, the models of these two tasks can be complementary to each other. On one hand, if an αNLG model can generate various kinds of hypotheses, it can provide more diversified instances for αNLI models to learn to discriminate different possibilities. On the other hand, if an αNLI model can effectively distinguish reasonable and unreasonable hypotheses, then the hypotheses generated from αNLG can leverage the discrimination signals from αNLI model to double-check to filter out confusing candidates.

Figure 1 shows the overall architecture of CAR. Specifically, CAR contains two critical component: 1) an ***abductive generator*** which conducts αNLG to generate diversified hypothesises; 2) an ***abductive discriminator*** which conducts αNLI to identify the plausible hypotheses from the generated results of abductive generator. Given a past observation O_1 and a future observation O_2, abductive generator will first sample several positive and negative candidate hypotheses, and then the abductive discriminator will be applied to determine the plausibility of these candidates. In this way, high-quality hypothesises from the abductive generator can be filtered out. Then these high-quality hypotheses are further used to refine the abductive discriminator to improve its ability to discriminate diversified instances. Then the procedure of plausibility estimation and abductive discriminator learning will be conducted iteratively, and therefore the filtered generated results from the abductive generator and the discrimination ability of the abductive discriminator can be jointly improved.

We conduct experiments on standard αNLI and αNLG benchmarks. Experiments show that CAR can significantly improve the generation quality of αNLG model and the discrimination ability of αNLI model and therefore achieve the state-of-the-art performance on both αNLI and αNLG leaderboard.

2 Related Works

Abductive Natural Language Inference Bhagavatula et al. (2020) first used BERT (Devlin et al., 2019) to compute the probability of each hypothesis then select the highest as plausible. Later, other pre-trained models including RoBERTa (Liu et al., 2019), DeBERTa (He et al., 2020) and UNIMO (Li et al., 2021b) were applied, achieving better results. Some works focused on knowledge enhancement (Paul and Frank, 2020; Banerjee et al., 2021; Du et al., 2021; Lourie et al., 2021; Zhou et al., 2021), while Paul and Frank (2021) generated the continuation of hypothesis to facilitate the inference. In addition to the simple binary classification method, Zhu et al. (2020) reconsidered αNLI as a ranking problem, and Li et al. (2021a) used the structural loss.

Abductive Natural Language Generation Bhagavatula et al. (2020) first fine-tuned GPT-2 (Radford et al., 2019), later T5 (Raffel et al., 2020) and UL2 (Tay et al., 2022) were available. Huang (2021) applied reinforcement learning, leveraging the α NLI model for supervised signal. Dong et al. (2021) applied attention modulation for better coherence of both observations. Zhou et al. (2022) introduced re-pretraining for better event understanding. Some works also tried to introduce additional knowledge for more plausible generation results (Bhagavatula et al., 2020; Ji et al., 2020;

Chakrabarty et al., 2021; Liu et al., 2022), while others focus on unsupervised learning (Qin et al., 2020; West et al., 2021; Qin et al., 2022).

3 Cycling Abductive Reasoner

Our proposed Cycle Abductive Reasoner consists of two components that interact with each other: Abductive Discriminator, which is initialized by αNLI task (Section 3.1), and Abductive Generator initialized by αNLG task (Section 3.2). In order to produce more coherent hypotheses, We use Abductive Discriminator to filter out less likely results among candidates sampled from the generator (Section 3.3). To further enhance the discriminator, we construct new diverse instances using both positive and negative generators to produce novel plausible and implausible hypotheses and examine them by the discriminator itself (Section 3.4).

3.1 Abductive Discriminator

The Abductive Discriminator is a binary classifier, usually an encoder-only pre-trained model like DeBERTa (He et al., 2021), that checks the coherence between the hypothesis H and observations O_1 and O_2. Given the two-segment input ($[CLS]$ O_1 $[SEP]$ H , O_2), where $[CLS]$ and $[SEP]$ are respectively classification and separation special token, the discriminator output the probability distribution of label y where $y = 1$ indicates plausibility and vice versa.

We fine-tune the discriminator on αNLI training set. Specifically, we decouple the original instance (O_1, O_2, H^+, H^-) into two, (O_1, O_2, H^+, $y = 1$) and (O_1, O_2, H^-, $y = 0$) respectively, where hypothesis H^+ is more plausible and H^- is less so.

3.2 Abductive Generator

In contrast, the abductive generator, a pre-trained model that supports text-infilling like encoder-decoder architecture T5(Raffel et al., 2020), generates hypotheses which happened in between the given observations. The input of the generator is "O_1 $[MaskToken]$ O_2", where for T5 $[MaskToken]$ is the special mask token "<extra_id_0>", and we use top-p sampling (Holtzman et al., 2020) with $p = 0.9$ to produce a large amount of hypothesis waiting to be filtered by the discriminator.

Since αNLG is in the form of text-infilling, we continually pre-train the generator G to maximize the log-likelihood of plausible hypothesis H^+ given O_1 and O_2. To augment the discriminator, we also use αNLI training set. However, generating positive instances alone may not be enough for data augmentation, so we also need to obtain a negative generator G^- simply by maximizing H^- on the instances (O_1, O_2, H^+, H^-).

3.3 Enhancing Generator with Discriminator

Although the generator can produce hypotheses related to the two observations at the token level, these hypotheses may not always be consistent with the observations at the sentence level. To overcome the imperfection of the generator, we re-evaluate the generations with the help of the Abductive Discriminator. Concretely, given the hypothesis candidates $C = \{h_1, h_2, ..., h_l\}$ sampled from Abductive Generator, the discriminator D throws out all less likely hypotheses that the probability of plausibility is lower than the threshold $\Theta = 0.9$, remaining plausible candidate set H^+:

$$H^+ = \{h | P_D(h|O_1, O_2) \geq \Theta, h \in C\} \quad (1)$$

Then, we choose the candidate with the highest log likelihood in terms of generator G as our final answer:

$$h^+ = \arg\max_{h \in H^+} logP_G(h|O_1, O_2) \quad (2)$$

3.4 Improving Discriminator with Generator

Many different events could happen after observation O_1, and at the same time trigger O_2 from happening. While in $\mathcal{ART}$ dataset, on average, there are 4.05 plausible and 9.37 implausible hypotheses for each pair of observations, this may not be the full picture. However, we could fill the gap and introduce more diversity by adding novel hypotheses produced by the aforementioned generation process. Specifically, for each pair of observations O_1, O_2 in the αNLI training set that appears k times, we use positive generator G with discriminator D to obtain positive hypotheses set H^+, and similarly use negative generator G^- for negative hypotheses set h^-. Then, we choose top-$2k$ most plausible hypotheses $h_1^+, ..., h_{2k}^+$ and top-$2k$ most implausible hypotheses $h_1 - +, ..., h_{2k}^-$ and construct new augmented instances $\{(O_1, O_2, h_i^+, y = 1) | 1 \leq i \leq 2k\} \cup \{(O_1, O_2, H^-, y = 0 | 1 \leq i \leq 2k)\}$. Training on the augmented dataset, we obtain the augmented discriminator that can determine with more confidence under more diverse situations.

Model	Baseline	CAR
BERT	72.3	**75.7**
Rainbow	79.5	-
RoBERTa	87.2	**88.6**
ALBERT	89.0	**89.9**
DeBERTa	92.6	**93.0**
Human performance	91.4	

Table 1: Accuracies of models before and after introducing our CAR on the test set of αNLI.

4 Experiments and Results

4.1 αNLI

Setup We use 304M DeBERTa-v3-large (He et al., 2021) as the primary model for αNLG filter and data-augmentation. To better study the quality of the augmented dataset, we also test 304M BERT-large-cased (Devlin et al., 2019), 304M RoBERTa-large (Liu et al., 2019) and 235M ALBERT-v2-xxlarge (Lan et al., 2020). We fine-tune these models both on the original αNLI training set and augmented dataset, with αNLI development set for hyper-parameter selection.

Results We report the results on Table 1, and human performance is copied from (Bhagavatula et al., 2020). The augmented DeBERTa model achieved the highest score, surpassing human performance also. All the models gain accuracy improvement on data augmentation, which shows that the augment dataset generated by our enhanced αNLG method is generally good in quality and yet novel compared with the existing instances.

4.2 αNLG

Setup We use 3B T5-v1_1-xl (Raffel et al., 2020) as our primary αNLG text generator, with augmented DeBERTa as the filter. We report the performance on αNLG test set. For comparison, we also copied the results of two previous SOTA models T5-11B (Khashabi et al., 2021) and 20B UL2 (Tay et al., 2022), and one unsupervised method DeLorean (Qin et al., 2020) from GENIE leaderboard[2], as strong baselines. Human performance is copied from (Bhagavatula et al., 2020). For ablation study, we remove the αNLI filter and only use top-p sampling ($p = 0.9$) or beam search ($beam\_size = 5$), and also replace the filter with the non-augmented one. Lastly, we try to select the final hypothesis by αNLI probability reranking.

[2] https://leaderboard.allenai.org/genie-anlg/submissions/public

Metric We use BLEU-4(Papineni et al., 2002) as our automatic metric. We also conduct human evaluation on 100 test instances, asking the evaluators to label the plausibility of hypothesis given the two observations and one generated hypothesis on a 4-point Likert scale, and each test instance was scored by three different evaluators.

Model	BLEU-4	Human
DeLorean (Unsup) (Qin et al., 2020)	1.84	45.13
T5-11B (Khashabi et al., 2021)	19.47	75.93
UL2 (Tay et al., 2022)	24.34	76.52
T5-xl + Beam-search	33.48	74.66
T5-xl + Sampling	33.69	73.84
T5-xl + Reranking	18.17	77.41
T5-xl + Filter	33.52	78.59
T5-xl + CAR (Ours)	**33.91**	**79.28**
Human performance	8.25	**96.03**

Table 2: Performance of models on the test set of αNLG, where only DeLorean is unsupervised method and others are fine-tuned. Human metrics refer to manual evaluation.

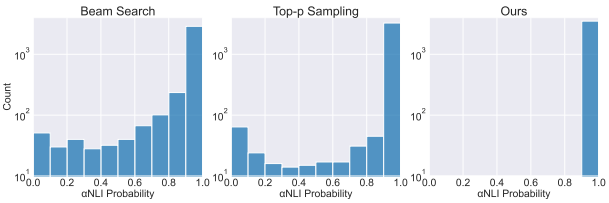

Figure 2: Distribution of αNLI Probability on αNLG test set, generated by beam search, top-p sampling and our method.

Results With the augmented filter, our method outperforms all other strong baselines on both automatic metric and human evaluation, achieving new State-of-the-art on αNLG. As illustrated in Figure 2, the discriminator throws out less coherent results that would have been chosen by beam search or top-p sampling. The ablation study also validates our introduction and augmentation of the filter and shows that reranking is more suitable for data augmentation.

5 Conclusion

We introduced Abductive Reasoning Cycle, leveraging the duality of its two task form. Experiments show that our approach creates new SOTA performance on both αNLI and αNLG tasks and enjoys a noticeable increase in the human evaluation of αNLG, which takes a step further toward human-level abductive reasoning ability.

Limitations

While we attempt to address the diversity issue of abductive reasoning by creating more novel hypotheses on observation pairs in trained dataset, when facing unseen observation pairs and their potential vast amount of hypotheses, our discriminator might still struggle, affecting the ability to precisely filter geneartion candidates, and consequently lagging behind human baseline in terms of human evaluation. One possible solution is to generate data on unseen observations for the discriminator so that it would be exposed to true diverse circumstances, but since it may not be robust enough when encountered with totally unseen data, we may not generate hypotheses with great confidence. We hope the future study can try to solve this dilemma.

Ethics Statement

This work involves no sensitive data, using public crowd-sourced dataset $\mathcal{ART}$ (Bhagavatula et al., 2020) only. This work aims to address the diversity issue of abductive reasoning, which makes it challenging to distinguish linguistically similar hypotheses in terms of plausibility, and hopes to inspire coherence and reasonable natural text generation.

References

Pratyay Banerjee, Swaroop Mishra, Kuntal Kumar Pal, Arindam Mitra, and Chitta Baral. 2021. Commonsense reasoning with implicit knowledge in natural language. In *3rd Conference on Automated Knowledge Base Construction*.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *ICLR*.

Tuhin Chakrabarty, Aadit Trivedi, and Smaranda Muresan. 2021. Implicit premise generation with discourse-aware commonsense knowledge models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6247–6252, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yue Dong, Chandra Bhagavatula, Ximing Lu, Jena D. Hwang, Antoine Bosselut, Jackie Chi Kit Cheung, and Yejin Choi. 2021. On-the-fly attention modulation for neural generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1261–1274, Online. Association for Computational Linguistics.

Li Du, Xiao Ding, Ting Liu, and Bing Qin. 2021. Learning event graph knowledge for abductive reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5181–5190, Online. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Hongru Huang. 2021. A reinforcement learning approach for abductive natural language generation. In *International Conference on Neural Information Processing*, pages 64–75. Springer.

Haozhe Ji, Pei Ke, Shaohan Huang, Furu Wei, Xiaoyan Zhu, and Minlie Huang. 2020. Language generation with multi-hop reasoning on commonsense knowledge graph. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 725–736, Online. Association for Computational Linguistics.

Daniel Khashabi, Gabriel Stanovsky, Jonathan Bragg, Nicholas Lourie, Jungo Kasai, Yejin Choi, Noah A Smith, and Daniel S Weld. 2021. Genie: A leaderboard for human-in-the-loop evaluation of text generation. *arXiv preprint arXiv:2101.06561*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Linhao Li, Ming Xu, Yongfeng Dong, Xin Li, Ao Wang, and Qinghua Hu. 2021a. Interactive model with structural loss for language-based abductive reasoning. *arXiv preprint arXiv:2112.00284*.

Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2021b. UNIMO: Towards unified-modal understanding and generation via cross-modal contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2592–2607, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ruibo Liu, Guoqing Zheng, Shashank Gupta, Radhika Gaonkar, Chongyang Gao, Soroush Vosoughi, Milad Shokouhi, and Ahmed Hassan Awadallah. 2022. Knowledge infused decoding. In *International Conference on Learning Representations*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Unicorn on rainbow: A universal commonsense reasoning model on a new multitask benchmark. *AAAI*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Debjit Paul and Anette Frank. 2020. Social commonsense reasoning with multi-head knowledge attention. In *In Findings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Debjit Paul and Anette Frank. 2021. Generating hypothetical events for abductive inference. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 67–77, Online. Association for Computational Linguistics.

Charles Sanders Peirce. 1974. *Collected papers of charles sanders peirce*, volume 5. Harvard University Press.

Lianhui Qin, Vered Shwartz, Peter West, Chandra Bhagavatula, Jena D. Hwang, Ronan Le Bras, Antoine Bosselut, and Yejin Choi. 2020. Back to the future: Unsupervised backprop-based decoding for counterfactual and abductive commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 794–805, Online. Association for Computational Linguistics.

Lianhui Qin, Sean Welleck, Daniel Khashabi, and Yejin Choi. 2022. Cold decoding: Energy-based constrained text generation with langevin dynamics. *arXiv preprint arXiv:2202.11705*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Yi Tay, Mostafa Dehghani, Vinh Q Tran, Xavier Garcia, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Neil Houlsby, and Donald Metzler. 2022. Unifying language learning paradigms. *arXiv preprint arXiv:2205.05131*.

Peter West, Ximing Lu, Ari Holtzman, Chandra Bhagavatula, Jena D. Hwang, and Yejin Choi. 2021. Reflective decoding: Beyond unidirectional generation with off-the-shelf language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1435–1450, Online. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wangchunshu Zhou, Dong-Ho Lee, Ravi Kiran Selvam, Seyeon Lee, and Xiang Ren. 2021. Pre-training text-to-text transformers for concept-centric common sense. In *International Conference on Learning Representations*.

Yucheng Zhou, Tao Shen, Xiubo Geng, Guodong Long, and Daxin Jiang. 2022. Claret: Pre-training a correlation-aware context-to-event transformer for event-centric generation and classification. *arXiv preprint arXiv:2203.02225*.

Yunchang Zhu, Liang Pang, Yanyan Lan, and Xueqi Cheng. 2020. L2r²: Leveraging ranking for abductive reasoning. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20.

A Training Details

A.1 αNLI

With 169,142 original αNLI training cases, we decouple the dataset into 71,915 positive examples and 166,137 negative examples, 238,052 in total. We generate 671,484 new instances, forming 909,536 augmented dataset. We train the model with the learning rate of $6e-6$ for 10 epochs with the warmup ratio of 0.06 and 96 batch size.

A.2 αNLG

We continually pre-train the models with the aforementioned 71,915 positive from αNLI training set. We train the model with the learning rate of $3e-5$ for 10 epochs with the warmup step of 50 and 192 batch size.

B Algorithm

Algorithm 1 Enhanced Generation

Require:
Abductive Generator (G)
Abductive Discriminator (D)
Observations O_1, O_2
Plausibility Probability Threshold Θ

function GENERATION(G, O_1, O_2)
  Sample candidate set C given O_1 and O_2
  $H^+ \leftarrow \emptyset$
  for c in C **do**
    $p \leftarrow P_D(c|O_1, O_2)$
    if G is Negative **then**
      // Filter by implausible probability
      $p \leftarrow 1 - p$
    end if
    if $p \geq \Theta$ **then**
      $H^+ \leftarrow H^+ \cup c$
    end if
  end for
  Rank H^+ by the log likelihood sum of G
  $h^+ \leftarrow$ hypothesis in H^+ with the highest score
  return h^+, H^+
end function

Algorithm 2 Augmentation for Discriminator

Require:
Positive Abductive Generator (G^+)
Negative Abductive Generator (G^-)
Abductive Discriminator (D)

function AUGMENTATION(O_1, O_2, k)
  $I \leftarrow \emptyset$
  // Generate positive hypotheses
  $\_, H^+ \leftarrow$ GENERATION (G^+, O_1, O_2)
  Rank H^+ by plausible probability
  $H^+ \leftarrow$ TOP-K(H^+, k)
  // Form positive instances
  for h^+ in H^+ **do**
    $I \leftarrow I \cup \{(O_1, h_i^+, O_2, y = 1)\}$
  end for
  // Generate negative hypotheses
  $\_, H^- \leftarrow$ GENERATION (G^-, O_1, O_2)
  Rank H^- by implausible probability
  $H^- \leftarrow$ TOP-K(H^-, k)
  // Form negative instances
  for h^+ in H^+ **do**
    $I \leftarrow I \cup \{(O_1, h_i^-, O_2, y = 0)\}$
  end for
  return I
end function

C Qualitative Analysis of αNLG

While Table 3 presents two additional automatic metrics, Table 4 present two concrete examples. On the first example, since in O_2 the mind reader told Jan the secrets of mind reading, it is highly possible that Jan asked him how he did it previously, which is exactly captured by our method, rather than asking if he could mind read, which beam search and top-p sampling fails and filter before data augmentation is being fooled. On the second example, our method generate less plausible hypothesis, for it's not convincing that grandma would be angry for being taught how to play given that O_1 already mentioned teaching, although yelling is not a polite behaviour, and the αNLI failed to discard this hypothesis. Also, not wanting to play Lumosity does not means you are not nice people, so beam search and filter before data augmentation also failed. Surprisingly, reranking yields a relatively good result, though not better than human written references.

Model	BLEU-4	ROUGE-L	BERTScore	Human
DeLorean (Qin et al., 2020)	1.84	19.67	88.54	45.13
T5-11B (Khashabi et al., 2021)	19.47	44.60	92.87	75.93
UL2 (Tay et al., 2022)	24.34	49.30	93.51	76.52
T5-xl + Beam-search	33.48	54.45	93.78	74.66
T5-xl + Sampling	33.69	54.45	93.74	73.84
T5-xl + Filter	33.52	54.41	93.75	78.59
T5-xl + rerank	18.17	43.43	92.47	77.41
T5-base + CAR (Ours)	14.99	40.30	92.11	60.25
T5-large + CAR (Ours)	23.22	46.62	92.89	70.33
T5-xl + CAR (Ours)	**33.91**	**54.58**	**93.81**	**79.28**
Human performance	8.25	30.40	-	**96.03**

Table 3: Performance of models on the test set of αNLG, with additional automatic metric of ROUGE-L (Lin, 2004) and BERTScore (Zhang* et al., 2020).

O_1	Jan went to the mind reader at the county fair.
H_{ours}	Jan asked the mind reader **how** he knew what she was thinking.
$H_{beam\ search}$	Jan wanted to know *if* the mind reader could read her mind.
$H_{top-p\ sampling}$	Jan asked the mind reader *if* he could read her mind.
$H_{no\ aug\ filter}$	Jan asked the mind reader *if* he could read her mind.
H_{rerank}	The mind reader asked her *what she wanted.*
H_{ref_1}	Jan asked **how** he knew things.
H_{ref_2}	Jan got her palms read.
H_{ref_3}	She asked **what** his secret was.
H_{ref_4}	She was surprised he could see her **skepticism**.
O_2	The mind reader told her that she had told him by her body language.

O_1	Lumosity is a fun game but it's not easy to teach to your Grandma.
H_{ours}	My grandma yelled at me for *trying to teach her*.
$H_{beam\ search}$	My grandma *didn't want to play lumosity*.
$H_{top-p\ sampling}$	My grandma yelled at me for *trying to teach her*.
$H_{no\ aug\ filter}$	My grandma *doesn't like to play games*.
H_{rerank}	We **got mad** at her for making an error.
H_{ref_1}	Grandma tried her best but we **laughed** at her.
H_{ref_2}	My husband and I **deleted** it off her phone.
H_{ref_3}	She started calling me names and **yelling**.
H_{ref_4}	We used to **confuse Grandma on purpose** so we could win.
O_2	We aren't very nice people.

Table 4: Generated examples of αNLG.