Improving Abstractive Summarization with Commonsense Knowledge

Pranav Ajit Nair
Indian Institute of Technology (BHU)
Varanasi
pranavajitnair.cse18@itbhu.ac.in

Anil Kumar Singh
Indian Institute of Technology (BHU)
Varanasi
aksingh.cse@iitbhu.ac.in

Abstract

Large scale pretrained models have demonstrated strong performances on several natural language generation and understanding benchmarks. However, introducing commonsense into them to generate more realistic text remains a challenge. Inspired from previous work on commonsense knowledge generation and generative commonsense reasoning, we introduce two methods to add commonsense reasoning skills and knowledge into abstractive summarization models. Both methods beat the baseline on ROUGE scores, demonstrating the superiority of our models over the baseline. Human evaluation results suggest that summaries generated by our methods are more realistic and have fewer commonsensical errors.

1 Introduction

Commonsense knowledge is crucial for understanding natural language. Several benchmarks have been developed to test commonsense reasoning skills in natural language processing systems. Most of these benchmarks such as HelaSwag (Zellers et al., 2019), WinoGrande (Sakaguchi et al., 2020) and CommonsenseQA (Talmor et al., 2019) formulate commonsense reasoning in the form of discriminative tasks. While significant progress has been made on such discriminative tasks, adding commonsense knowledge and reasoning skills into natural language generation still remains a challenge.

Recent work has addressed this challenge to some extent. Tasks such as Abductive commonsense generation (Bhagavatula et al., 2020) test commonsense reasoning ability of large pretrained models and demonstrated that the performance on such tasks is far behind human performance. Yang et al. (2019) added commonsense knowledge into sequence to sequence models via external knowledge graphs. The recently proposed CommonGen (Lin et al., 2020) benchmark also tests commonsense reasoning skills of models via language generation.

Inspired from recent work on commonsense knowledge generation tasks we propose two training techniques to add commonsense knowledge to abstractive summarization models. While conducting experiments on the baseline models we observe that the generated summaries lack commonsense reasoning in them, i.e the models are unable to capture basic commonsense knowledge while generating summaries. For example, the model generated the phrase ‘researchers, known as the excess uric acid’. This phrase does not comply with our general commonsensical knowledge. The model is unable to derive the correct relation between a ‘researcher’ and ‘uric acid’.

Our approaches try to rectify the above problem in abstractive summarization models. Our first approach is similar to Yang et al. (2019) wherein we use external commonsense knowledge to supply additional reasoning power to the model. By doing so we are able to add basic commonsense knowledge about the concepts into our model. Our second approach is a training-finetuning based approach wherein we finetune the model to generate summaries conditioned only on the concepts in the document and not the entire document. In this way we are able to relate different parts of the document since the model has to implicitly extract relations between these concepts to generate a summary. We assume that the concepts extracted from the document in some way represent every part of the document. Both our approaches solve the aforementioned problem to some extend and also improve the performance of the summarization models in terms of evaluation metric such as ROUGE. We also try a few other techniques which we mention in later sections.

Proceedings of the Student Research Workshop associated with RANLP-2021, pages 135–143,
held online, Sep 1–3, 2021.
https://doi.org/10.26615/issn.2603-2821.2021_019

2 Related Work

Abstractive summarization aims at capturing the entire essence of a document within a small number of words. There has been great progress made in the recent past where better and better architectures have been proposed to solve the problem. See et al. (2017) used pointer generator networks to copy words from the input document. Duan et al. (2019) augmented the Transformer architecture with a contrastive attention mechanism to ignore the irrelevant parts of the document. Zhang et al. (2020) pretrained a Transformer model for summarization. Our work unlike previous work tries to add commonsense knowledge into abstractive summarization models with minimum change to the training objective and model architecture.

Addition of commonsense knowledge and commonsense reasoning abilities into machine learning models has been a topic of interest in recent years. Several benchmarks and tasks have been established to evaluate commonsense reasoning skills. These tasks are based on temporal commonsense reasoning (Zhou et al., 2020), commonsense knowledge about physical entities (Bisk et al., 2020), world knowledge based commonsense reasoning (Talmor et al., 2019), abductive commonsense inference (Bhagavatula et al., 2020), human emotion (Rashkin et al., 2018), coreference resolution (Sakaguchi et al., 2020), sentence completion (Zellers et al., 2019) etc.

Another set of commonsense reasoning oriented tasks are based on natural language generation. One such task is abductive commonsense generation (Bhagavatula et al., 2020) wherein given a set of observations the model has to generate a hypothesis which explains the second observation based on the first observation. Another commonsense language generation benchmark is language generation conditioned on concepts (Lin et al., 2020) wherein the model has to generate a plausible natural language sentence given a set of concepts.

Recent work has also been focused on adding commonsense knowledge into sequence to sequence models. Yang et al. (2019) added external commonsense knowledge extracted from knowledge graphs into topic-to-essay generation models. They use ConceptNet (Speer et al., 2017) to extract concepts and relations. Zhou et al. (2021) added commonsense knowledge to pretrained language models, their approach is inspired by the CommonGen benchmark (Lin et al., 2020).

For abstractive summarization Amplayo et al. (2018) developed an entity linking system to extract linked entities knowledge graphs. These linked entities carry commonsense knowledge since they are linked to a knowledge graph and are used to generate a representation for the topic of the summary. This representation is used to guide the decoder. Feng et al. (2021) developed a dialogue heterogeneous graph network using utterances and commonsense knowledge for abstractive dialogue summarization.

Our approach draws ideas from these works and applies them to abstractive summarization. Firstly we explore methods to add external commonsense knowledge into models. We use Numberbatch embeddings pretrained on ConceptNet to add commonsense into our model similar to Yang et al. (2019). Our next approach is based on generating summaries only from document concepts. The problem setting is similar to that of CommonGen (Lin et al., 2020) where the model needs to implicitly reason between the concepts to generate the summary. We also try a fusion of these approaches. Lastly we explore pretraining-finetuning based approaches wherein we pretrain our model on other commonsense reasoning tasks and then use it train our summarization model.

3 Method

In this section we present our two methods to add commonsense knowledge and reasoning skills into abstractive summarization models. All our models are based on Transformers (Vaswani et al., 2017). Let the document be represented by $x = (x_1, x_2, ..., x_n)$ where n is the length of the document and let the summary be represented by $y = (y_1, y_2, ..., y_m)$ where m is the length of the summary. Let the Transformer model encoder be represented by enc and let the Transformer decoder be represented dec, at the t^{th} time step the model generates the t^{th} word of the summary as:

$$P(y_t|y_1, ..., y_{t-1}, x) = \text{softmax}(dec(y_1, ..., y_{t-1}, enc(x))W) \quad (1)$$

where $W \in \mathbb{R}^{d \times |V|}$ is the embedding matrix and V is the Vocabulary.

3.1 Extracting Concepts

Both our approaches rely on concepts extracted from the documents. We use nouns and verbs as concepts to train our model. Let the Concepts be represented by $C = \{c_1, c_2, ..., c_K\}$ where K is total number of concepts in the document (not necessarily distinct). In our ablation studies we also use adjectives along with verbs and nouns as input concepts to the model.

3.2 Adding External Commonsense Knowledge

We use Numberbatch embeddings of the extracted concepts in our model. Since Numberbatch embeddings are trained on ConceptNet they carry world knowledge in them. The encoder in our model is the same as that of the Transformer encoder. The decoder has been modified. The self attention part remains the same in the decoder, for the cross attention part the decoder attends to the concept embeddings along with the encoder output. A linear transformation is applied to the concept embeddings before cross attention to match the Transformer model dimension. The model equation can be represented as:

$$\begin{aligned}&P(y_t|y_1, ..., y_{t-1}, C, x) = \\ &\quad \text{softmax}(dec(y_1, ..., y_{t-1}, enc(x), \\ &\qquad W_1\text{NB}(C))W) \quad (2)\end{aligned}$$

where NB represents the Numberbatch embeddings derived from ConceptNet and W_1 is a linear transformation applied to concept embeddings before cross attention to match the Transformer model dimension. The cross attention equations can be represented as:

$$\begin{aligned}Q &= W_Q enc(x) \\ K &= [W_K enc(x); W_1\text{NB}(C)] \\ V &= [W_V enc(x); W_1\text{NB}(C)] \quad (3) \\ Attention(Q, K, V) &= \text{softmax}(\frac{QK^T}{\sqrt{d}})V\end{aligned}$$

where $W_Q \in \mathbb{R}^{d \times d}; W_K \in \mathbb{R}^{d \times d}; W_V \in \mathbb{R}^{d \times d}$ and ; represents the concatenation operation. Note that the above equation has been written for a single head, in practice we use mutli-head attention. Our model is trained with cross entropy loss. We call this approach NBEmbed. Next we present our second approach to add commonsense knowledge and reasoning capabilities.

3.3 Training Only With Concepts

First we train a Transformer model with the entire document until convergence. We further train the model wherein it is either given the entire document or only the document concepts to generate the summary. The model equation can be represented as:

$$p \sim \text{Bernoulli}(\phi) \quad (4)$$

$$\begin{aligned}&P(y_t|y_1, ..., y_{t-1}, C, x) = \\ &\quad \text{softmax}(dec(y_1, ..., y_{t-1}, p \times enc(C) \\ &\qquad + (1 - p) \times enc(x))W) \quad (5)\end{aligned}$$

We do not change the relative order of the concepts as encountered in the document. We call this concept only training approach ConTra. In the next section we provide the experimental details and our results.

4 Experiments

We use the CNN/Daily Mail dataset for all our experiments. All our models are Transformer based, we use the Transformer base model with 6 encoder layers, 6 decoder layers, 8 attention heads, 512 is the hidden and embedding dimension and 2048 dimensional feed forward layer. We use the linear warm-up followed by square root decay schedule proposed by Vaswani et al. (2017). We train our baseline for 200K iterations with a batch size of 32 and we employ early stopping based on validation loss. We use beam search decoding in all our experiments with a beam size of 4, we also employ a length penalty of 1.0. We use label smoothing with $\alpha = 0.1$. While generating the summaries we do not let the model repeat trigrams (Paulus et al., 2018). We use Sentencepiece tokenizer in all our experiments. All the experiments are performed using PyTorch on a single NVIDIA Tesla V100 GPU.

4.1 Results with addition of external commonsense knowledge

The results on validation set can be found in Table 1. As one can clearly see from Table 1 that our proposed technique outperforms the baseline in all the three ROUGE scores. Since the Numberbatch embeddings are 300 dimensional vectors the linear mapping used is of size 300×512. We observe that given the external knowledge our model is better able to relate between the concepts. The generated summaries are more realistic which we confirm via human evaluation. Unless otherwise specified the Numberbatch Embeddings of the concepts are

Model variant	ROUGE-1	ROUGE-2	ROUGE-L
Baseline	30.60	9.97	27.63
NBEmbed	**31.29**	10.11	28.48
ConTra	31.07	**10.34**	28.45
ConTra+NBEmbed	30.71	10.15	**28.50**

Table 1: Results on the validation set obtained for our best performing configurations.

Model variant	ROUGE-1	ROUGE-2	ROUGE-L
Baseline	29.74	9.43	27.11
NBEmbed	30.13	9.60	27.41
ConTra	30.15	**9.80**	27.35
ConTra+NBEmbed	**30.41**	9.77	**27.99**

Table 2: Results on the test set obtained for our best performing configurations.

given for both training and testing only to the decoder. Results on the test set can be found in Table 2.

4.2 Results by training with only Concepts

The results on validation set can be found in Table 1. Clearly our model outperforms the baseline. The most optimal setting is to train the baseline for full convergence and train the model further for 200K iterations with $\phi = 0.5$. The other settings can be found in the ablation study section. We observe that since the model is given only concepts to generate the summary in the later stages of training it implicitly learns to reason among the concepts and makes less commonsensical errors which we confirm via human evaluation. We also try out another version of our model where we combine the concept only training approach with external commonsense knowledge addition approach. Although there is no significant gain over the two proposed models' performances, it still beats the baseline. Unless otherwise specified the model is always given the entire document for testing. Results on the test set can be found in Table 2.

4.3 Ablation Study

We conduct ablation study to try out various configurations, which may or may not produce better results. We present these results in Table 3 and Table 4. For the part where we use external commonsense knowledge in the decoder, we also try a version where this knowledge is provided in the encoder self attention section (NBEmbed+encNBEmbed), the results are presented in Table 3. We see that the model performance significantly decreases, probably because providing a knowledge distribution different from that of the encoder for self-attention adds noise which reduces the informativeness of the encoder self-attention.

We also try a version of our model where the external knowledge is provided only during the validation or the testing phase (NBEmbed-only_val). We see that the performance is similar to the baseline. The model is not able to utilize this knowledge since it had not seen it during training, these results are also presented in Table 3 for the validation data.

For our concept only training style we experiment with multiple training and finetuning styles. Firstly we only train the model on concepts, that is the model has never seen the entire document while being trained, for testing and validation we try two settings where we either provide the model with the entire document (Concept_only+Document) or we provide the concepts only as done in training (Concept_only+Concept). In both the cases the performance is significantly below the baseline. Since the model is never exposed to the entire document it does not learn the grammatical aspects of summary generation and is thus unable to generate fluent summaries. Even if the entire document is provided for testing the model does not know how to utilize most of the grammatical information in the document. These results are presented in Table 4 for validation data.

In our next setting we pretrain the model with the entire document and finetune only using the concepts, that is, in the finetuing stage the model is not given entire documents (ConTra-Document). Even in this case the performance drops slightly below the baseline. Probably because, since only concepts are being used for finetuning the model, there is less stability in this stage of learning. These

Model variant	ROUGE-1	ROUGE-2	ROUGE-L
NBEmbed+encNBEmbed	15.49	4.40	16.26
NBEmbed-only_val	30.30	9.80	27.77
NBEmbed+Adjective	31.07	9.93	27.78

Table 3: Results for variants of our first approach on the validation set.

Model variant	ROUGE-1	ROUGE-2	ROUGE-L
Concept_only+Concept	19.56	3.20	17.83
Concept_only+Document	24.45	4.38	22.91
ConTra-Document	29.76	9.52	27.20
ConTra-Concept_only+Concept	25.48	5.66	23.36
ConTra-Concept_only+Concept+Document	24.85	5.27	23.10
ConTra+Adjective	30.57	9.89	27.72

Table 4: Results for variants of our second approach on the validation data set.

Model variant	ROUGE-1	ROUGE-2	ROUGE-L
Pretraining on Abductive natural language generation	5.72	0.44	6.39

Table 5: Results for the pretraining approach from section 4.4 on the validation set

results for validation set are presented in Table 4.

We try two other settings where the model is pretrained with concepts and finetuned either with only documents (ConTra-Concept_only+Concept) or with a mix of both documents and concepts (ConTra-Concept_only+Concept+Document). In both the cases the model's performance is below that of the baseline. Probably, since majority of the training is done with only concepts the model is unable to learn grammatical aspects of generating entire summaries. These results for the validation set are presented in Table 4.

Next we add adjectives in the document to our concept set as well, we call these models NBEmbed+Adjective and ConTra+Adjective. The results on the validation set are shown in Table 3 and Table 4 respectively. As we can see from the results there is no increase in the performance, since adjectives in isolation, when disconnected from the noun they modify, cannot truly express the noun's property.

4.4 Another Variant

Here we try to add commonsense knowledge into the model via pretraining on a dataset requiring commonsense reasoning abilities. For this we use the abductive natural language generation dataset proposed by Bhagavatula et al. (2020). In this dataset the model is required to generate a hypothesis explaining two observations. Since this task requires understanding and reasoning about the two observations to generate a hypothesis, in some ways it might be similar to summarization where the summary has to be generated based on the given document although the amount of reasoning required is far less. Thus we pretrain the model on abductive natural language generation dataset and then further train it on the CNN/Daily-Mail dataset. Results on the validation set can be found in Table 5. As one can see in Table 5 the model performance drops significantly, probably because the small size of the dataset and difficulty of the task adds too much noise into the model parameters from which the model is unable to recover in the latter stages of training.

4.5 Qualitative Analysis

To analyze the generated summaries qualitatively we present summaries generated by NBEmbed in Table 6. In the first sample the phrase 'west sussex coast's nephew ned rocknroll' generated by the baseline is commonsensically incorrect since a person cannot be a location's nephew. Such errors are rarely made by NBEmbed. In the second example the phrase 'the 35-year-old from erik compton' is commonsensically incorrect since 'the 35-year-old' cannot be located in a human being ('erik compton'). Again, NBEmbed makes no such errors. Similarly in the thrid example the baseline generates the phrase 'playing for arsenal's official website', which does not make sense, since a person cannot

Article: kate winslet faces a planning battle with natural england over plans to build sea defences . kate winslet is to face a battle with natural england over plans to build a 550ft-long wall to protect her £3.25 million beachside home from flooding and erosion. the multi bafta-winning actress, 39, wants to install a boulder and gravel seawall along the west sussex coast to protect her property. but natural england has recommended the planning application be refused - citing concerns it may result in the ' direct loss' of natural habitat designated for rare and vulnerable birds, trees and wetlands. the academy award-winning actress bought the eight bedroom house in west sussex in 2013. she lives there with her husband, richard branson' s nephew ned rocknroll, and her three children .(Truncated) **NBEmbed**: multi bafta-winning actress, richard branson' s nephew ned rocknroll has recommended the planning application to install a boulder and gravel seawall along her £3.25 million beachside home from flooding . **Baseline**: multi bafta-winning actress, 39, wants to install a boulder and gravel seawall along the **west sussex coast' s nephew ned rocknroll** . but natural habitat designated for her husband richard branson' grade ii home from flooding in 2013 .
Article: we' re suckers for athletes who play with a lot of heart. in erik compton' s case, make that plural: hearts. he' ll tee off this week at the masters on his third. the 35-year-old veteran would rather be known for his golf than what he' s had to overcome to earn a spot in the field. but compton isn' t shy about sharing his story, either. erik compton shares a joke with jim furyk during a practice round at augusta ahead of the masters . a first heart transplant at 12, then a second in 2008 after a major heart attack - what doctors call a ' widow-maker' - while driving home from a lesson with longtime coach and close pal charlie delucca. he' s been trying make up for lost time ever since. ' it' s hard, but it' s also a great thing,' compton said on monday after a practice round at augusta national. ' there' s two sides to me. (Truncated) **NBEmbed**: erik compton isn' t shy about sharing his story at the masters in erik recifeca, georgia on monday and close pal charlie deluc-maker's case to earn a joke with jim furyk during practice round . **Baseline**: **the 35-year-old from erik compton** shares a joke with jim furyk on his third . the old u.s in 2008 after an practice round at augusta ahead of masters' s two penalty for national park am i, who is set to become first major heart transplant by four other european championship west ham united states he says: 'the my mind sometimes you'
Article: former italy forward antonio di natale has described alexis sanchez as the best strike partner he' s ever had and insisted the arsenal forward is better than neymar. di natale, who won 42 caps for italy between 2002 and 2012, spent five years playing alongside sanchez at udinese, who he still captains at the age of 37. azzurri legends francesco totti and alessandro del piero are just two of the names di natale has played alongside but he told arsenal' s official website that sanchez tops the lot . (Truncated) **NBEmbed**: former italy forward antonio di natale has been in all competitions during the age of 37. azzurri legends francesco totti and alessandro del piero' s ever-mate, who won 42 caps at serie a side udinese . **Baseline**: antonio di natale has played alongside former team-mate alexis sanchez at the age of five years **playing for arsenal' s official website** in a storming form with 20 goals totti and alessandro del piero .

Table 6: Summaries generated by NBEmbed from test set. Commonsensical errors made by the baseline are highlighted.

Article: nairobi, kenya (cnn)university of nairobi students were terrified sunday morning when they heard explosions – caused by a faulty electrical cable – and believed it was a terror attack, the school said. students on the kikuyu campus stampeded down the halls of the kimberly dormitory, and some jumped from its fifth floor, the university said. hundreds were injured and were taken to hospitals. one person died, according to the school. the confusion and panic came less than two weeks after al-shabaab slaughtered 147 people at a college in garissa, kenya . (Truncated) **ConTra**: one person was injured, a university official says . the confusion and panic came less than two weeks after al-shabaab slaughtered 147 people at college in garissa. kenyan teachers were admitted to kenyatta national hospital' s students have been discharged as many are slated for surgery . **Baseline**: **a faulty electrical cable was a terror-related school** in west kenya on sunday morning . at least 63 people were injured, the university of nairobi students have been discharged after an lillian lepos. normal power supply will resume its ceo' s african retcing incident this week .
Article: cradling your stomach with a hot water bottle might seem like the norm at ' that time of the month' . however, if you are doubling over with debilitating pain, it could actually be a sign of something much worse. while most ' normal' period pain can be fixed with ibuprofen or a over-the-counter anti-inflammatory, a doctor tells daily mail australia that more severe symptoms such as nausea, back or leg pain or bleeding at unexpected times of the month can all point to a more serious condition - endrometriosis. dr lara briden explains that endometriosis is a condition where bits of the uterus lining grow in other places outside of the uterus such as the ovaries, bladder or intestines. painful periods: 90 per cent of women will experience period pain in their life time . (Truncated) **ConTra**: dr lara briden explains that endometriosis is a condition where bits of the uterus such as an anti-inflammatory or severe symptoms can be suffering from your run . **Baseline**: dr lara briden says the wants pain or bleeding at ' normal period, bladder-counter anti. 90 per cent of **menstrtriosis** is a condition where bits in all point to more serious health .
Article: green party activists have been told to dress in ' mainstream' fashion while knocking on doors in a bid to win over sceptical voters. a manual for the party' s supporters urges them to appear ' level headed' and ' agreeable' – and even encourages them to compliment people' s homes. the advice, which has been distributed among green campaigners in london, also provides stock answers to ease voters' concerns about their radical plans to dismantle the army, legalise drugs and pay everybody £72 a week no matter how rich they are. scroll down for video . (Truncated) **ConTra**: a manual for the party' s supporters urges them to dress in , even encourages their radical plans and pay everybody £72 a week no matter how rich they are . **Baseline**: a manual for the party' s only mp caroline lucas at a fracking protest in support of green campaigning **$1-called on doors**, west sussex . it comes amid growing scrutiny to more than 50,000 and ukip - unlike her key housing policies .

Table 7: Summaries generated by ConTra from test set. Commonsensical errors made by the baseline are highlighted.

be playing football for a website. Again such errors are not committed by the NBEmbed model.

We also, perform a similar qualitative study for ConTra. Samples generated by ConTra are presented in Table 7. In the first sample the phrase ‘a faulty electrical cable was a terror-related school’ does not make sense commonsensically since a ‘faulty electrical cable’ cannot be a school. Such errors are not committed by ConTra. In the second sample the baseline uses ‘menstrtriosis’ instead of ‘endometriosis’, which is erroneous since there is no condition called ‘menstrtriosis’. ConTra gets it right ans does not make such mistakes. Again in the third sample the phrase ‘$1-called on doors’ generated by the baseline does not appear anywhere, in the entire article. Such hallucinations are avoided by ConTra.

4.6 Human Evaluation

We randomly sampled 50 examples from test set for human evaluation. All evaluators were engineering undergraduates with professional working proficiency in English. Each evaluator was presented with the original document, the summary generated by the baseline and the summary generated by NBEmbed objective model, and each summary was evaluated by exactly one evaluator. 40 evaluators were involved in the process. We asked the evaluators to rate the summaries based on two factors: 1) Commonsensical errors made by the model and 2) Overall quality of the summary (i.e relevance to highly informative parts of the document and linguistic quality). The evaluators had to pick the better summary or flag them as equally bad or good with respect to the two evaluation criteria. In 36% of the cases the human evaluators found the summaries generated by NBEmbed to have made fewer commonsensical errors than the baseline, whereas only in 10% of the cases the evaluators found the baseline to be better. In 36% percent of the cases the evaluators preferred NBEmbed summaries over the baseline, whereas only in 24% of the cases the baseline was preferred.

We conducted a similar human evaluation for ConTra. In 40% of the cases the human evaluators found the summaries generated by ConTra to have made lesser commonsensical errors than the baseline, whereas only in 28% of the cases the evaluators found the baseline to be better. In 46% percent of the cases the evaluators preferred ConTra summaries over the baseline, whereas only in 38% of the cases the baseline was preferred.

Thus from human evaluation results we concluded that both our approaches make less commonsensical errors and generate more realistic summaries.

5 Conclusion

In this work we proposed two techniques to add commonsense knowledge to abstractive summarization models. We observe that the baseline is sometimes unable to generate summaries that comply with the general notion of commonsense. Both our approaches solve this issue to a certain extend. Our first method is based on adding external commonsense knowledge into the model, and our second method is based on training the model with only concepts present in the document rather than using the entire document. We show that our models outperform the baseline on the evaluation metric. We also experimented with several variants of our proposed approaches. Our human evaluation results suggest that the summaries generated by our models are indeed more realistic. The application of our methods on state-of-the-art summarization systems should further improve the results. One limitation to be addressed is to formalize the notion of ‘commonsensical soundness’ and come up with an automatic evaluation metric to measure it. In future we would work on this limitation. We would also experiment with other techniques and devise better methods for adding commonsense knowledge into abstractive summarization models and extend our methods to other natural language generation tasks such as machine translation.

Acknowledgements

We would like to thank the anonymous reviewers for their insightful comments and suggestions. The support and the resources provided by the PARAM Shivay Facility under the National Supercomputing Mission, Government of India at the Indian Institute of Technology (BHU), Varanasi, are gratefully acknowledged. We also thank Manav Jain for helping with human evaluation and Samyak Jain for comments on early drafts of the paper.

References

Reinald Kim Amplayo, Seonjae Lim, and Seung won Hwang. 2018. Entity commonsense representation for neural abstractive summarization. In *Proceedings of the 2018 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Association for Computational Linguistics.

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 *International Conference on Learning Representations*.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7432–7439.

Xiangyu Duan, Hongfei Yu, Mingming Yin, Min Zhang, Weihua Luo, and Yue Zhang. 2019. Contrastive attention mechanism for abstractive sentence summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics.

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. Incorporating commonsense knowledge into abstractive dialogue summarization via heterogeneous graph networks. In *Chinese Computational Linguistics - 20th China National Conference, CCL 2021, Hohhot, China, August 13-15, 2021, Proceedings*, volume 12869 of *Lecture Notes in Computer Science*, pages 127–142. Springer.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *International Conference on Learning Representations*.

Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. 2018. Modeling naive psychology of characters in simple commonsense stories. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8732–8740.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Pengcheng Yang, Lei Li, Fuli Luo, Tianyu Liu, and Xu Sun. 2019. Enhancing topic-to-essay generation with external commonsense knowledge. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2002–2012.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Ben Zhou, Qiang Ning, Daniel Khashabi, and Dan Roth. 2020. Temporal common sense acquisition with minimal supervision. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

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