Generating Conclusions from Medical RCT Papers

Te-Wei Shieh[1,3], Yung-Sung Chuang[2]

1. School of Medicine, National Taiwan University, Taipei, Taiwan.
2. Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan.
3. Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan.

Introduction

Today, a myriad of data is generated in the clinical research community every day. It would be helpful for machines to understand and analyze them for human readers. In this study, we evaluate the effectiveness of sequence-to-sequence models on generating conclusions from medical randomized controlled trial papers. First, we transformed a medical text classification task into a conclusion generating task. Then, a standard RNN-with-attention model and a more advanced GPT-2 model were trained on the generation task. We further discussed the performance and limitations of the current models.

Material & Methods

Dataset

PubMed 200k RCT: A typical RCT often constitutes of two randomized groups of patients receiving either the "intervention" (new treatment) or "control" (conventional treatment). Then, a statistical analysis is done after the experiments to determine whether the intervention has a significant effect (i.e. actually making patients better or worse).

This dataset was originally constructed for sequential short text classification. We concatenated the "background", "objective" and "results" sections of the abstracts and asked the model to generate the "conclusions". The first 5 words of the first conclusion sentence were appended in the input as a hint.

Our task is similar but not identical to summarization. Writing conclusions is more than selecting important information and paraphrasing. In order to generate correct and fluent conclusions, the model must identify the objectives of the experiment and interpret the results. Therefore, generating comprehensible conclusions requires both sufficient natural language understanding and generation capabilities.

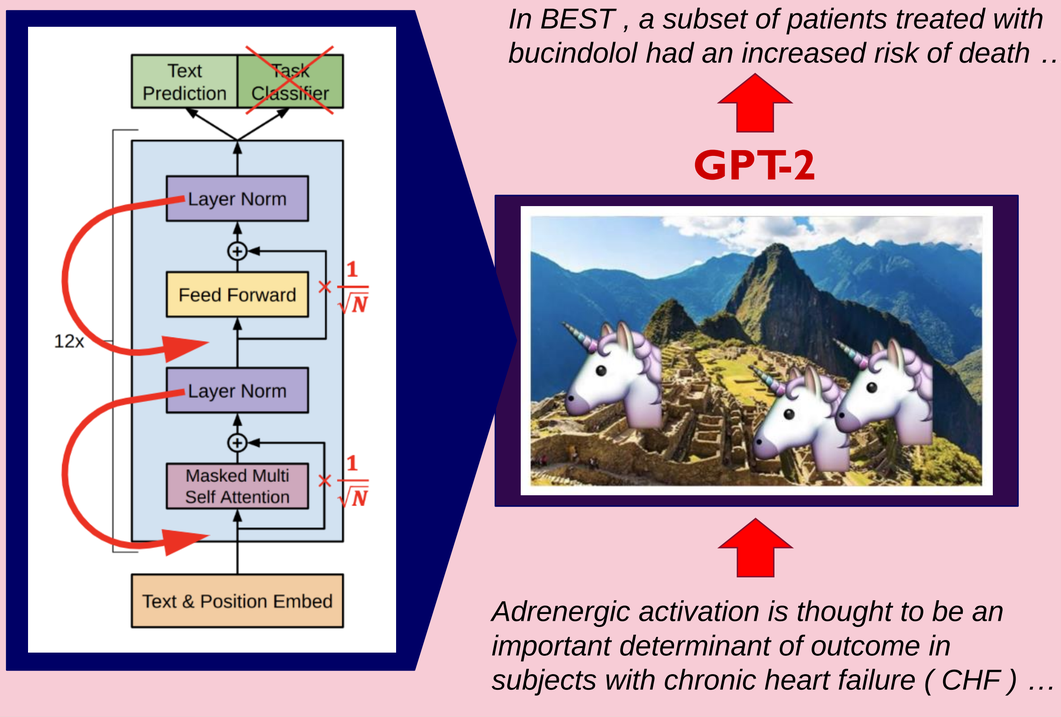
Model and Experiments

Bidirectional RNN (GRU) with Attention

- Standard machine translation / sequence-to-sequence model
- Ran the experiments for 50 epochs and selected the model with the best validation BLEU score for manual evaluation.

Open AI GPT-2 (**G**enerative **P**re-**T**raining)

- Transformer-based language model capable of multitask learning.
- Our modification allowed GPT-2 to generate new conclusions while being fully aware of the given context and hint
- Fine tuned for 5 epochs



Results

- Results are based on the validation set of PubMed 200k RCT.
- Manual evaluation was done on the first 50 examples generated by each model.

Model	Val Perplexity	Val BLEU
Seq2Seq + Attention	20.1910	10.9250
Fine Tuned GPT-2	9.7345	10.8496

Model	Success	False Positive	False Negative
Seq2Seq + Attention	34%	6%	6%
Fine Tuned GPT-2	46%	8%	8%

Results

Source : To assess the impact of intraoperative distal vasal flushing during no-scalpel vasectomy on hastening the sperm clearance from the vas deferens and subsequent postvasectomy time to azoospermia . … A significantly higher proportion of patients in group B were azoospermic at 4 and 8 weeks when compared with group A … There was no statistically significant difference in the mean age , operative time , and number of ejaculations between the 2 groups . Conclusion : Distal vasal flushing during no
Target : Distal vasal flushing with 30 mL of sterile water may shorten the time to azoospermia for between 20 % and 30 % of vasectomy patients . Thus , this procedure may be an option for some who choose the vasal flush to avoid the long duration of postvasectomy alternate contraceptive methods .
Prediction [Seq2Seq] : distal <unk> flushing with 30 ml of flushing during no - scalpel vasectomy does not reduce the number of patients .
Prediction [GPT-2] : Distal vasal flushing during no-scalpel vasectomy is associated with a higher incidence of hastening the sperm clearance from the vas deferens and subsequent postvasectomy time to azoospermia.

Target : Laparoscopic port-site skin closure with octylcyanoacrylate is rapid and effective . The method yields cost savings and a decrease in operative time of more than 9 minutes per case .
Prediction [Seq2Seq] : laparoscopic port - site skin closure with subcuticular sutures is associated with a lower cost of skin closure than suture closure .
Prediction [GPT-2] : Laparoscopic port closure with octylcyanoacrylate is a safe and effective technique for closing surgical skin incisions.

Target : In BEST , a subset of patients treated with bucindolol had an increased risk of death as the result of sympatholysis , which compromised the efficacy of this third-generation beta-blocker .
Prediction [Seq2Seq] : in best , a subset of patients with chf , bucindolol did not increase the risk of death or death or chf hospitalization . the benefit of the drug - specific agent was not associated with an increased risk of death or death or chf hospitalization .
Prediction [GPT-2] : In BEST, a subset of subjects with CHF who received bucindolol had a significantly higher risk of death or death plus CHF hospitalization.

Target : The combination therapy of doxycycline/sulphadoxine/pyrimethamine appeared to be significantly effective in the treatment of patients with chloroquine resistant falciparum malaria without causing any serious side effects . …
Prediction [Seq2Seq] : the combination therapy of doxycycline / artesunate / pyrimethamine is effective in preventing the prevention of <unk> in the treatment of <unk> .
Prediction [GPT-2] : The combination therapy of doxycycline and sulphadoxine/pyrimethamine is effective in the treatment of Sudanese patients infected by chloroquine resistant falciparum malaria.

Conclusion

- We defined a new conclusion generation task from the PubMed 200k RCT dataset.
- Human evaluation showed that the quality of the conclusions generated by our GPT-2 model is generally better than that of the baseline Seq2Seq model.
- Our GPT-2 model demonstrated better paraphrasing capabilities. With byte-pair encoding (BPE), it can overcome OOV problems without the need for a pointer mechanism.

Acknowledgment

We thank Prof. Yun-Nung Chen and Shang-Yu Su for their insightful discussions.

References

1. Franck Dernoncourt, Ji Young Lee. PubMed 200k RCT: a Dataset for Sequential Sentence Classification in Medical Abstracts. IJCNLP, 2017.
2. Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. ICLR, 2015.
3. Radford and Wu et al. Language Models are Unsupervised Multitask Learners.

國立臺灣大學資訊工程學系