Towards neural abstractive clinical trial text summarization with sequence to sequence models

Celia Cintas*, William Ogallo†, Aisha Walcott‡, Sekou L Remy §, Victor Akinwande ¶ and Samuel Osebe‖
IBM Research Africa. Nairobi, Kenya
Email: *celia.cintas@ibm.com, †william.ogallo@ibm.com, ‡awalcott@ke.ibm.com, §sekou@ke.ibm.com,
¶victor.akinwande1@ibm.com, ‖sosebe@ke.ibm.com

***Abstract*—The recruitment stage in clinical trials is key in ensuring enrollment of a large and diverse number of participants. Recent trends in clinical trials recruitment strategies have leveraged social media, mobile, and web-based platforms to advertise trials to a broader and more diverse set of potential participants. We develop a method to improve clinical trials enrollment rates through novel models of communication that provide accurate and unbiased information about the clinical trials and provide awareness to target participants. The contributions of this paper are two-fold. First we propose a model to generate abstractive summaries for clinical trials based on sequence to sequence networks with attention policies. Second, we present a preliminary evaluation of the model in terms of learning, vocabulary development, choices of attention policies, and summarization outputs. Finally, we generate a dataset consisting of multi-sentence clinical trials summaries to be used for bench-marking and in future work.**

I. Introduction

An estimated 11% of sites in multicenter, global clinical trials fail to enroll patients, and almost 40% fail to meet their initial recruitment targets [1]. Such difficulties in participant recruitment can result in extended recruitment time, increased costs for recruitment, as well as inadequate statistical power if enrollment targets are not met [2]. Newer methods for running clinical trials, such as site-less or eClinical trials, have started to take advantage of technology, internet, social media, and telemedicine to address the recruitment and enrollment problems, and to automate data collection, improve control over intervention content and format, as well as reduce costs [3], [4]. Our aim is to improve clinical trial enrollment rates through novel models of communication which present accurate and unbiased information about the clinical trials to target participants. Generating useful information, and presenting it in a readily consumable format from a large, complex document without direct human intervention is a challenging task. The amount of information and domain-specific medical terminology makes it even more challenging. This paper presents preliminary work on the development of a model that takes as input long free-text descriptions about a clinical trial and generate a short human-interpretable description that can be used to improve the engagement and recruitment of potential trial participants.

II. Training the Seq2Seq Model

In our approach, we leverage a bidirectional Recurrent Neural Network [6] to implement the model, and to train it, we created a novel dataset. The dataset is comprised of sets of pairs containing short and long descriptions of clinical trials.

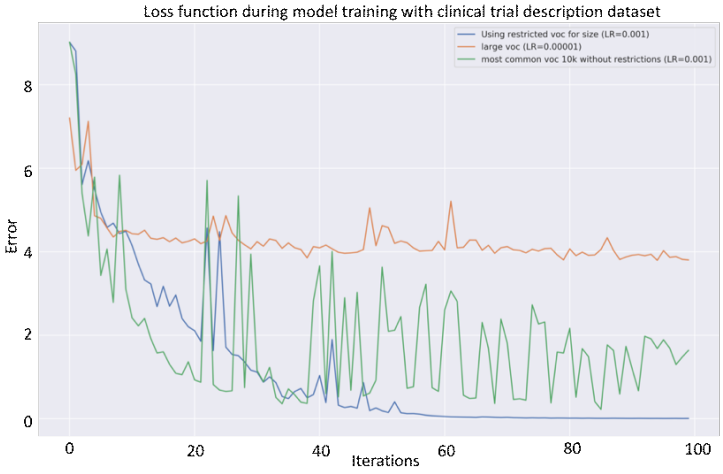

Fig. 1: Loss curves produced by the training set applied to our seq2seq network. These curves capture three combinations of learning rates and vocabulary policies.

A. Dataset

We started with $286,000$ records of clinical trials downloaded from a publicly accessible database [1]. We filtered the records which had the pair (S, D), where S is a sequence of sentences in the summary, and D is a sequence of sentences of the detailed clinical trial description, and ensured that $|S| < |D|$. To the resulting elements, we applied classic NLP pre-processing steps such as the removal of stop words, removal of contractions, stemming [7], and grouping words with the WordNet lemmatizer [8]. The average number of tokens per summary S is 85, and the average number of tokens in full descriptions D is 323. The resulting clinical trial dataset is then composed of $120,459$ pairs, which was split into a training subset containing $84,321$ pairs (70%), a validation set of $12,045$ (10%) and a testing set of $24,093$ pairs (20% of the total dataset) that were selected with a random permutation cross-validation iterator.

B. Vocabulary Building

As for the vocabulary, initially we had $173,761$ words, plus the tokens for padding ($< PAD >$), end of sentence ($< EOS >$), start of sentence ($< SOS >$), and unknown terms ($< UNK >$). To speedup training and reduce memory use, we filtered the maximum word length, generating a new

[1] https://clinicaltrials.gov/

978-1-5386-9138-0/19/$31.00 ©2019 IEEE

Setup	ROUGE-1			ROUGE-2			ROUGE-L		
	F	P	R	F	P	R	F_{lcs}	P_{lcs}	R_{lcs}
full voc. with Attention model	0.377	0.353	0.415	0.149	0.135	0.168	0.328	0.317	0.3777
batch voc. with Attention model	**0.404**	0.559	0.327	**0.150**	0.182	0.131	**0.338**	0.533	0.309
batch voc. without Attention model	**0.247**	0.246	0.256	**0.043**	0.047	0.041	**0.208**	0.214	0.218

TABLE I: ROUGE-1, ROUGE-2 and ROUGE-L precision, recall and F-score on test sets with different vocabulary configuration and attention model. In red the lowest F-score correspond to Seq2seq models without any Attention mechanism.

D_1	this trial is being conducted to determine the safety and efficacy of injecting autologous adipose derived cells into the penis of men with erectile dysfunction $< UNK >< EOS >$ subjects will be randomized into a treatment and control arm $< EOS >$
S_1	$====$ the purpose of this study is to evaluate the safety and efficacy of adipose derived cells for the treatment of erectile dysfunction symptoms $< EOS >$
G_1	$>>>>$ the purpose of this study is to evaluate the safety and tolerability of the drug the term of term of erectile $< UNK >$ in patients $< EOS >$
D_2	the purpose of this single center study is to test whether smart anemia manager algorithm improves hemoglobin stability compared to standard anemia management protocol $< EOS >$ study site kidney disease program dialysis facility university of louisville louisville ky $< EOS >$
S_2	$====$ first clinical evaluation of smart anemia manager algorithm $< EOS >$
G_2	$>>>>$ the clinical evaluation of smart manager algorithm $< EOS >$
D_3	to improve the quality of the surgery of the pouch created in cu and fap patients the following information will be collected patients characteristics operative data postoperative outcome and functional postoperative recovery $< EOS >$
S_3	the purpose of this registry is to collect information about patients in which a pouch has been created to improve in the future the quality of the surgery of the pouch $< EOS >$
G_3	the purpose of this study is to determine which is the best surgery $< EOS >$

TABLE II: Sentences generated by the model trained with scheduled sampling [5] and restricted vocabulary per batch. Where D_i is the full description, S_i is the ground truth summary and G_i the generated summary by the model. In red, word repetition, in orange words missing in the generated summary G_i. The F-score over LCS for the G_i shown in the Table are$F_{lcs}^{G_1} = 0.593$, $F_{lcs}^{G_2} = 0.817$ and $F_{lcs}^{G_3} = 0.435$.

vocabulary of $73,557$ unique words. We next analyzed the generated vocabularies based on the frequency of words and their distribution in the clinical trials dataset. In the training stage, the vocabulary was restricted to the batch tokens, both in the source and target text. This drastically reduced the convergence time of the network. Fig. 1 illustrates how different policies of vocabulary generation affect the training error curves. In Table I, the ROUGE scores for restricted batch vocabulary versus full vocabulary show that the first policy improves the metric notably.

C. Training setup considerations

The major drawback of seq2seq with a recurrent neural network architecture is that training the network is very resource intensive, requiring days or even weeks of GPU time to converge. This makes the hyper-parameter search activity excessively time consuming. Due to this, our initial setup of hyper parameters were based on the work of [9]. They provide extensive exploration of Neural Machine Translation architectures and hyper-parameters, that can be adapted to summarization problems. The embedding dimension was 512, following [9], and the RNN cell variant was GRU. We used a batch-size of 128 elements per iteration and randomly shuffled the training data. We used dropout for regularization [10] and gradient clipping [11]. Based on [12] and the analysis shown in Section II-B, we used a small vocabulary setup (10k words) and restricted by batch, reducing the training iterations and the converge time. During training, we applied maximum likelihood sampling. This resulted in the generation of coherent grammar but inaccurate information in the clinical trial summary. To reduce this error, we used scheduled sampling [5].

III. Analysis and Results

The comparative ROUGE metrics with and without attentions mechanism can be seen in Table I. The *ROUGE* [13] metrics are averaged over all the elements. We confirm that the use of attention mechanism is imperative to generate more accurate clinical trials summaries compared to plain seq2seq models. Based on qualitative analysis, the use of schedule sampling [5] during training generates more accurate information in the summary. Several output examples of sentences generated by the model trained with schedule sampling are provided in Table II. The loss functions depicting the training error of the network with different vocabulary policies are shown in Fig. 1. Clearly, the use of the restricted vocabulary per batch improves both the ROUGE metrics and the content accuracy as shown in Table I and Table II. Also, training converges more quickly in this case, as can be seen in the Fig. 1.

IV. Conclusions and Future Work

To improve clinical trials recruitment rates, from the perspective of communication issues, it is critical that potential participants are made aware of the clinical trials that are planning to or are in the process of recruitment. Potential participants should be provided with correct and summarized information about the clinical trials they are about to participate in. Our analysis shows the feasibility of using seq2seq with attention policies for automatic feature extraction and summarization of biomedical texts. The proposed solution presents promising results, but more experiments need to be done. These include, train the model over longer sentences to extend its capabilities to larger documents and compare the similarity and correctness of the summaries generated by the model versus human experts.

REFERENCES

[1] I. B. Hirsch, J. Martinez, E. R. Dorsey, G. Finken, A. Fleming, C. Gropp, P. Home, D. I. Kaufer, and S. Papapetropoulos, "Incorporating Site-less Clinical Trials Into Drug Development: A Framework for Action," *Clin Ther*, 2017.

[2] B. G. Sully, S. A. Julious, and J. Nicholl, "A reinvestigation of recruitment to randomised, controlled, multicenter trials: A review of trials funded by two UK funding agencies," *Trials*, 2013.

[3] E. Murray, Z. Khadjesari, I. R. White, E. Kalaitzaki, C. Godfrey, J. McCambridge, S. G. Thompson, and P. Wallace, "Methodological challenges in online trials," *Journal of Medical Internet Research*, 2009.

[4] J. A. Anguera, J. T. Jordan, D. Castaneda, A. Gazzaley, and P. A. Areán, "Conducting a fully mobile and randomised clinical trial for depression: Access, engagement and expense," *BMJ Innovations*, 2016.

[5] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, "Scheduled sampling for sequence prediction with recurrent neural networks," *CoRR*, vol. abs/1506.03099, 2015. [Online]. Available: http://arxiv.org/abs/1506.03099

[6] T. Shi, Y. Keneshloo, N. Ramakrishnan, and C. K. Reddy, "Neural Abstractive Text Summarization with Sequence-to-Sequence Models," dec 2018. [Online]. Available: https://arxiv.org/abs/1812.02303

[7] C. D. Paice, "Another stemmer," *ACM SIGIR Forum 24.3*, pp. 56–61, 1990.

[8] C. Fellbaum, "WordNet: An Electronic Lexical Database(Language, Speech, and Communication)," *The MIT Press*, 1998.

[9] D. Britz, A. Goldie, M. Luong, and Q. V. Le, "Massive exploration of neural machine translation architectures," *CoRR*, vol. abs/1703.03906, 2017. [Online]. Available: http://arxiv.org/abs/1703.03906

[10] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," pp. 1–9, 2014. [Online]. Available: http://arxiv.org/abs/1409.3215

[11] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[12] R. Nallapati, B. Xiang, and B. Zhou, "Sequence-to-Sequence RNNs for Text Summarization," *CoRR*, vol. abs/1602.06023, 2016. [Online]. Available: http://arxiv.org/abs/1602.06023

[13] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries," *Annual Meeting of the Association for Computational Linguistics*, 2004.