What's next

R&D

Using deep learning to generate medical reports

Doctors are using more and more of their time on writing documentation, which is time away from interacting with patients. There have been attempts to automate the process of creating reports based on patient-doctor conversations, but the existing systems are complex and require manual annotation. Sequence-to- sequence models could in principle be used for translating a conversation into a report. We compared how well different architectures are suited for summarizing orthopedic encounters.

Seppo Enarvi
Posted July 10, 2020

A large portion of a physician's time goes to documentation. There has been an increase in

medical documentation requirements and this burden has been identified as one of the main contributing factors for physician burnout. An important part of this documentation is a report that is produced after every patient encounter. Automatic speech recognition (ASR) technology helps doctors by letting them dictate reports instead of typing them on a keyboard. However, if the content of the report is already discussed during the patient visit, writing or dictating it is seen as a redundant task, and could in principle be automated.

Earlier attempts at automatic report creation have focused on extracting clinical information from patient-doctor conversations, which could eventually be formatted into a text report using templates. Unfortunately, such pipelines are complex and require manual annotation of clinical information in the training data. Annotation is often too expensive to scale such systems to growing amounts of data.

Sequence-to-sequence models have been applied to various natural language tasks, such as machine translation and summarization. In the paper "Generating Medical Reports from Patient-Doctor Conversations using Sequence-to-Sequence Models", we study how well they could be applied to medical report generation. We compare network architectures that are based on the traditional RNN encoder-decoder model, and newer Transformer architectures.

We transcribe the patient-doctor conversations into text using ASR. A sequence-to-sequence model is trained to summarize the text conversation into a report. We incorporate enhancements in the RNN and Transformer summarization models in a novel way to mitigate their limitations. For RNN models we use a hierarchical encoder following Cohan et al. (2018) that processes chunks of the input sequence in parallel to speed up training. To facilitate copying words from the conversation, we incorporate a pointing mechanism, inspired by See et al. (2017). Our Transformer model is depicted in Figure 1.

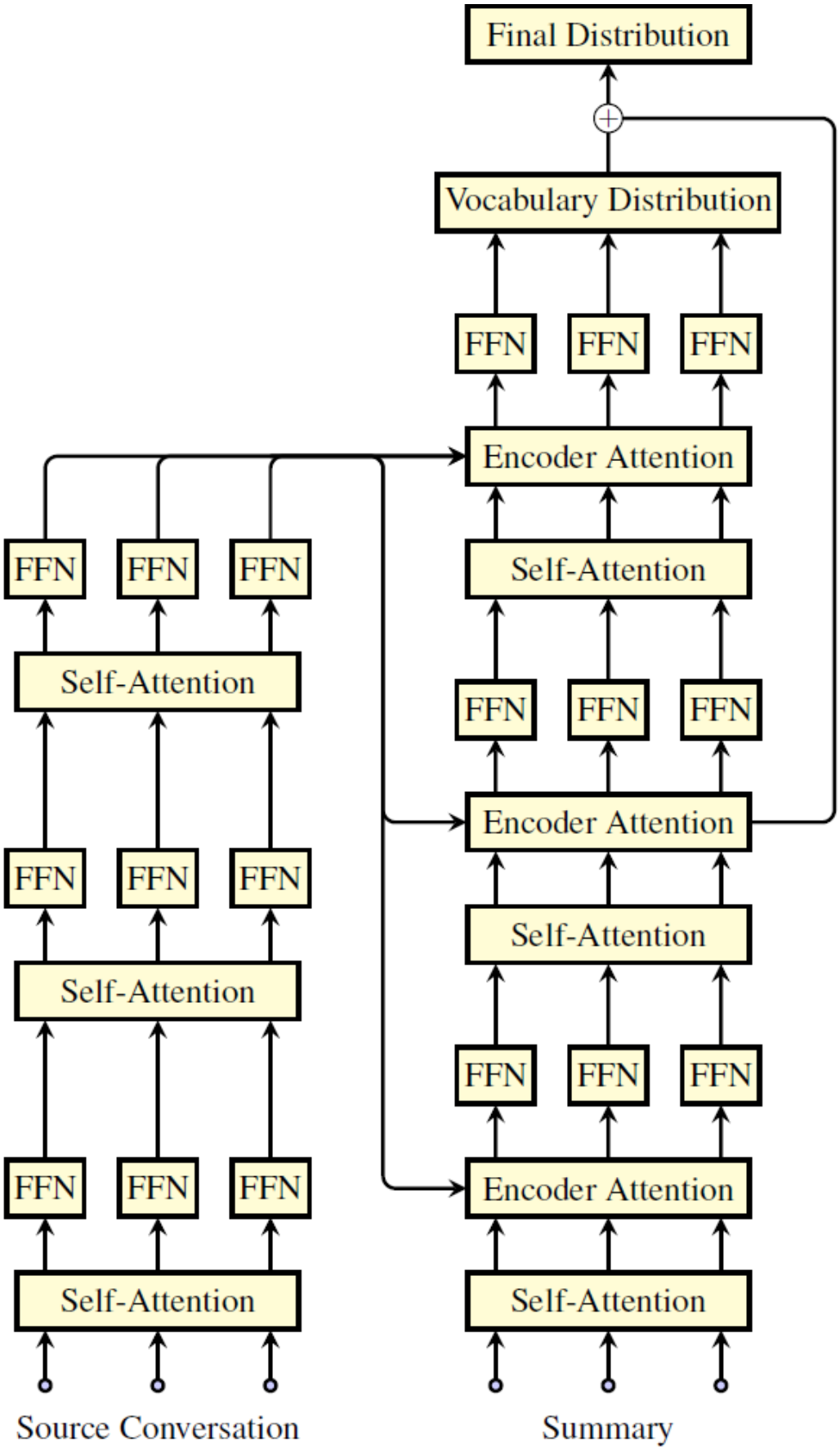
Final Distribution
Vocabulary Distribution
FFN
FFN
FFN
Encoder Attention
Self-Attention
FFN
FFN
FFN
Encoder Attention
Self-Attention
FFN
FFN
FFN
Encoder Attention
Self-Attention
FFN
FFN
FFN
Self-Attention
FFN
FFN
FFN
Self-Attention
FFN
FFN
FFN
Self-Attention
Source Conversation
Summary

Figure 1. We add a pointing mechanism to the Transformer model to facilitate copying of input words. One attention head of the penultimate decoder layer is taken as a probability distribution over the source tokens and interpolated with the normal output distribution. This enables copying of source tokens, even if they are not in the vocabulary.
We apply the models on a large corpus of conversation-report pairs from orthopedic patient visits. Models are trained a maximum of one week on 8 Nvidia v100 GPUs in Azure. During this time, the Transformer models have reached convergence, but the RNN models mostly have not. The hierarchical RNN encoder is faster to train, progressing three times as many training steps as the normal RNN model. In a practical scenario with limited computational resources, the hierarchical RNN model is clearly advantageous. Transformer-based models achieved significantly better accuracy, however, while taking less than three days to train. The pointing mechanism further improved performance of both RNN and Transformer models in most cases.

Figure 2 shows an example conversation and a report created by a Transformer pointer-generator model. The conversation is a simulation of a real patient encounter. Several facts are omitted from the generated report. We've also observed information repetition and hallucinations that are not grounded in the conversation. The model has captured most of the information from the conversation correctly and formulated it in the appropriate jargon and style of a medical report. For example, high blood pressure is referenced as hypertension and gallbladder removal is summarized as cholecystectomy.

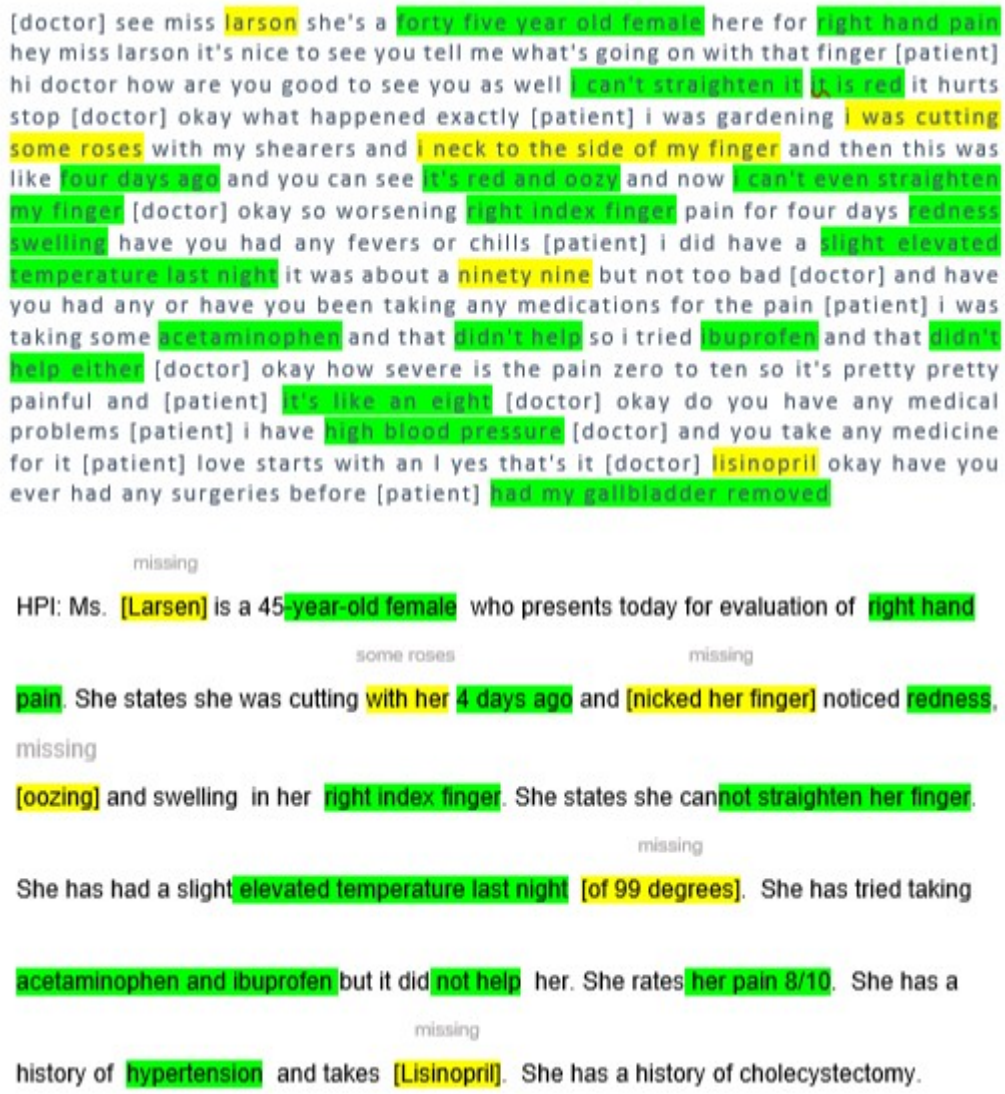

Figure 2. Artificial conversation (above) and model output (below). Although not a real patient encounter, this example demonstrates summarization capability and mistakes made by a Transformer pointer-generator model. Facts that the model has captured correctly are highlighted with green, and missing or incorrect information is highlighted with yellow.

Our results indicate that sequence-to-sequence models, in particular Transformer, are able to generate relatively fluent and factually correct reports from transcribed conversations between a doctor and a patient. The models are in many cases able to formulate information in a manner appropriate for a medical report, instead of just extracting word sequences from the conversation. However, there's more work to do as these generated reports also exhibit errors common to such models. Further analysis is needed to better assess report quality and contrast with pipelined approaches.

The aforementioned paper was presented at the First Workshop on Natural Language Processing for Medical Conversations, which was part of the 58th ACL 2020 conference, the premier conference of the field of computational linguistics, and was awarded Best

Paper .

Acknowledgments

This paper was co-authored by Seppo Enarvi, Marilisa Amoia, Miguel Del-Agua Teba, Brian Delaney, Frank Diehl, Stefan Hahn, Kristina Harris, Liam McGrath, Yue Pan, Joel Pinto, Luca Rubini, Miguel Ruiz, Gagandeep Singh, Fabian Stemmer, Weiyi Sun, Paul Vozila, Thomas Lin, and Ranjani Ramamurthy.

Tags: Deep Learning

About Seppo Enarvi

Seppo Enarvi is a Senior Research Scientist in the Nuance DAX team. His interests lie in deep learning, especially sequence modeling. He worked on language modeling at Aalto University prior to joining Nuance in 2017, and obtained a Ph.D. soon after. He enjoys coding and has a long history in software development.

View all posts by Seppo Enarvi