Abstractive Summarization: A deep reinforced learning approach

Shreedhar Kodate

Mid-term MTech Project Report

Abstract

Precis/summary writing is a very interesting task for humans. After collecting all kinds of information, it is being compressed and stored in abstractively summarized way, let's say "small text". This small text can now be distributed as quick source of important information. It can also be expanded back into loads of data based on certain and different requirements. Neural sequence-to-sequence models with various techniques viz. recurrent networks, pointer-generator, coverage vector, intra-attention, multi-sentence summaries have improved the state-of-the-art for Abstractive Summarization. In this project we intend to twist and tweak various existing models to understand how these techniques would act as different dimensions of summarization and if there are any overlaps. We would also like to find out if there is a "dominating technique" that can take the state-of-the-art to the next level. We will be proposing a new model BKA which is intuitionally similar to what humans do as part of their day-to-day learning.

1 Introduction

Summarization

Precis/summary writing is inherently an important task in our daily lives. We come across loads of textual information and at the back of our mind, we are continuously summarizing the data for easy future reference. It is important to note that we don't summarize randomly nor we summarize everything of it, we store just enough information to be able to regain all the original data back (by querying various sources of information). Basically, Text summarization is of two types: Extractive and Abstractive. Extractive method means to select (based on some heuristics and parameters) and arrange sentences from the input document(s) to form smaller paragraph(s). Abstractive summarization is similar to what humans do. This method should incorporate sophisticated abilities that are crucial to high-quality summaries, such as paraphrasing, generalization, or the incorporation of real-world knowledge, which is possible only in an abstractive framework[3].

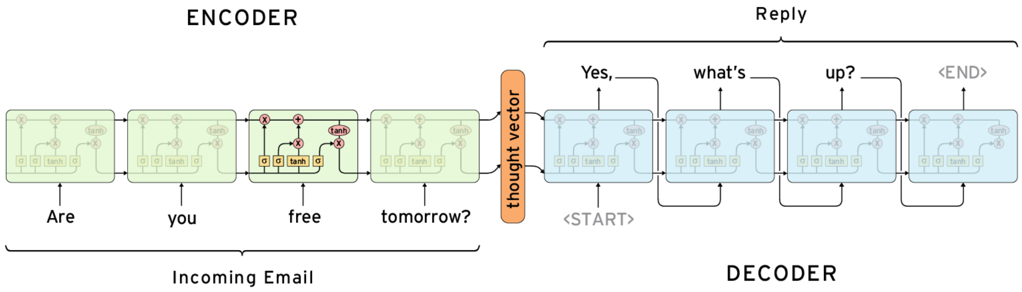
Abstractive summarization is a kind of Generative model which do not depend entirely on the input sentences. A sentence may indirectly mean something else. But we need to model the sentences/phrases in such a way that the latent meaning is revealed to decode later. That is to generate sentences from scratch. Such Generative models which are somewhat based on Machine translation techniques can be used to translate the input document(s) to output summaries. These models can also refer back to the contextual entity information like intricate details of an event(refered to earlier) being described. Sequence to sequence Deep learning architectures are specially suited for generating text and there's a hope of rapid progress of research in this area.

Abstractive models also need to keep track of what has been summarized, what remains to summarize, what is the limit for the size of the output, uniqueness of the words used, salience of the topics covered and yet to be covered, etc. Attentional model plays a very important role to keep a comprehensive track of the input document (Shadowing). Summary is nothing but a small talk with the person reading it. So, a summary's intention plays an intrinsic role. We also need to capture the dynamics of the intention process to create more human-like summaries. Recurrent neural networks have proven to be very effective in modelling and recreating Attention and Intention in the world of textual data.

Figure-1[1] explains a basic RNN based encoder and decoder scheme.

[1] Source: http://www.wildml.com/2016/04/deep-learning-for-chatbots-part-1-introduction/

Figure 1: RNN based encoder-decoder with a magic thought box



State-of-the-art

Attentional Recurrent neural network (RNN) based encoder-decoder models have proven to perform good on short input and output sequences, but are repetitive and filled with incoherent phrases when faced with longer documents and summaries. Hence, intra-attention has been introduced in such a model and is combined with the global sequence prediction training of Reinforcement Learning (RL) by using policy learning algorithm to generate more natural summaries. The resulting summaries became more readable and they could also tackle problems like ‘exposure-bias’[1]. RL is basically a way of training an agent to gain maximum reward in a certain environment by monitoring it’s own behavior. It is not hard to perceive that we should be able to train a agent to generate good summaries(rewarding) after conceptualizing the document content (interaction with environment). Before moving to other research, we would like to highlight that no one metric is good enough to measure the quality of an automatically generated summary. Although the best metric would be to let humans decide but that is not scalable and well-quantified metric. There are human biases, feelings, mood and various other perplexities associated with such a metric. There are standard metrics available for Evaluation of the generated summary like ROGUE, BLEU, NIST etc. But we would like to express our concern to construct a hybrid metric that is fast, precise and reliable.

Novel models based on Attentional Encoder-Decoder RNNs that address critical problems in summarization that are not adequately modeled by the basic architecture, such as modeling keywords, capturing the hierarchy of sentence-to- word structure, and emitting words that are rare or unseen at training time and show that they achieve state-of-the-art performance on two different corpora[2]. A hybrid pointer-generator network that can copy words from the source text via pointing, which aids accurate reproduction of information, while retaining the ability to produce novel words through the generator. Combined with the coverage vector to keep track of what has been summarized, which discourages repetition, they outperform the current abstractive state-of-the-art by at least 2 ROUGE points[3]. A structurally simple local attention-based model that generates each word of the summary conditioned on the input sentence. It can easily be trained end-to-end and scales to a large amount of training data[4]. Emphasizing the importance of the attention mechanism in the paper with results as following. The Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data[5]. An ensemble of models sets a new record by improving perplexity from 41.0 down to 23.7[6]. Ptr-Nets not only improve over sequence-to-sequence with input attention, but also allows to generalize to variable size output dictionaries[7]. The performance of a seq2seq model can be improved when two binary vectors are used to track the decoding stack in transition-based parsing, and multi-layer attention is introduced to capture multiple word dependencies in partial trees [8]. SummaRuNNer though producing extractive type of summaries have incorporated some abstract features like information content, salience and novelty in their method[9]. Summarizing by transforming a source semantic graph to a summary graph using Abstract meaning representations (AMR) and parsing it to generate output[10].

Dataset

We will use the CNN/Daily Mail dataset which contains online news articles (781 tokens on average) paired with multi-sentence summaries (3.75 sentences or 56 tokens on average). The data has 287,226 training pairs, 13,368 validation pairs and 11,490 test pairs. We will operate directly on the original text (or non-anonymized version of the data), which we believe is the favorable problem to solve because it requires no pre-processing [3].

2 BKA Model

Summaries need to be sensible enough to incorporate both linguistic and physical context. Like in long conversations, people remember what has been said and what information exchange has already happened. Following a recent common approach in textual modelling, we can try to embed/convert a conversation into vectors, but doing that for long dialogs is challenging. Similarly, vectorizing long sentences with the latent meaning captured is a difficult task. RNNs can be used to compute conditional probabilities of phrase pairs to learn semantically and syntactically meaningful representation of linguistic phrases to generate meaningful sentences in decoder stage. Salience of a word/phrase/sentence can be measured as the quality of being prominent in a document(s). This can be used to highlight the usefulness of the generated summaries.

After studying the various arguments as presented in previous sections, we did a thought experiment to understand what is actually the process of communication between two individuals. Because, all the knowledge boils down to just vibrational signals transmitted and received via a medium. So, we have come up with (not so new) idea of representing an entity's (sender/reciver) brain consisting of 3 main blocks: Background-Knowledge-Application(BKA). In the first block, the entity has some preconceptions which account to the base belief systems, assumptions, facts and theories. All this is essential part of our memory and intelligence. Also, we know that not everything is ground truth and may change in future, hence this Background block has a wonderful property called as plasticity.

Whenever this block comes into contact of new data, roadblock, information, the entity enters into the second block of operation as Knowledge where the entity can derive many kinds of inferences, judgement, understanding of the data in hand and try to visualize and extract some information. Later these extracts/excerpts are formalized into the Background block for further usage as required. We take a moment to emphasize that all the methods discussed in the state-of-the-art section capture one or more (but not all) of the varied capabilities of human brain. We need a novel blend of methods that can capture the information processing that human brain does. How to capture the essence of the input being read in O(1) time? is one of the question, we'll try to answer.

The third and the final block is Applications. Whatever the entity has assimilated, it is put to use to derive new results and conclusions of unseen/anticipated data. Applications is not only limited to act/react when put in a certain situation, but to prepare itself for challenges long before they present themselves. We'll basically try to implement such a model and experiment with the existing techniques in NLP. We choose Deep Reinforcement Learning, because of the inherent requirement of the tools needed to solve the problem of abstractive summarization viz. sequence-to-sequence, networking between the internal components, knowledge representation, actions and rewards, continuous play between an entity and environment.

Figure-2[2] shows how an architecture for deep visual attention. We shall implement a similar deep attentional and intentional architecture to capture important charecteristics like salience and readability.

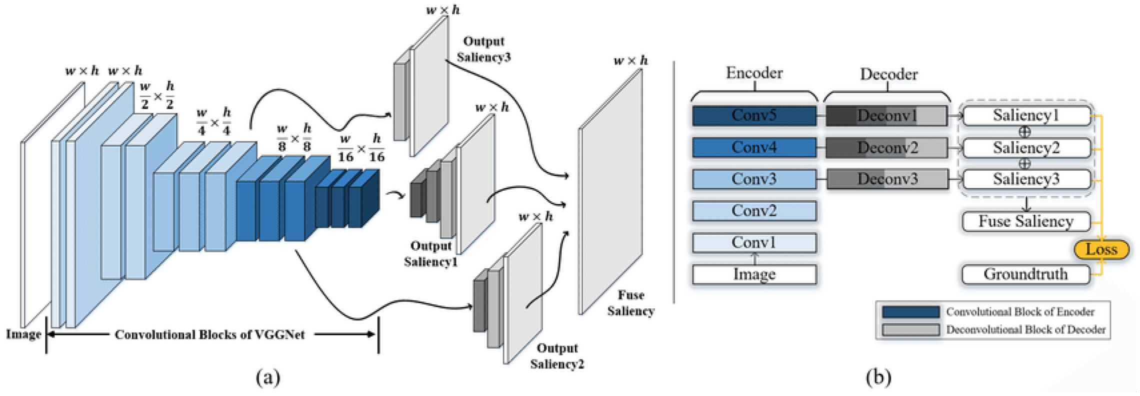
3 Conclusions and Future Work

We will be experimenting with various techniques that we have come across in automatic summary generation tasks to find out which works best. Then, we'll combine these techniques to create a dominating technique to produce results beyond the state-of-the-art. Once, our summarizer is empowered to generate good quality, human-like summaries, we would also proceed to work on following significant applications that have a potential to change the way we look at the world.
[Daily short news] In today's fast paced world, we would like to be updated as quickly as possible with the current affairs with facts and figures as per our individual needs. Inshorts (smartphone app) is a very good example of what we are trying to do but is manual driven. The company employs a team that manually curates and provides the 60 word summary for each feed that is divided into categories such as national, business, sports, and technology [Wikipedia]. If such summaries were generated by machines with a good human readability score, all the content of the internet can be made available like a small hand book.

[Summarizing for different groups] People can be divided into various kinds of groups eg. educational qualification, geographical locations, bodily age, etc. Each group has a different kind of characteristics, intelligence and knowledge. Ideally speaking, we

[2] Source:https://www.researchgate.net/publication/316779608_Deep_Visual_Attention_Prediction

Figure 2: VGGNet based encoder-decoder model for Deep Visual Attention



should present a message in different ways for different groups. This method of delivery of knowledge has various flaws (intelligent audience get bored with basic concepts, children might face knowledge gaps, interests of a group are not met). Once we have a strong abstractive summarizer, the same content can be summarized in various ways to suit the needs of a group. With this, education system can be revolutionized by gradual mastery based teaching.

[Auto-Knowledge Staircase] With an abstractive summarizer acting on the gigantic internet, each and every subject matter can be made available at the touch of a button. The huge amounts of data can be summarized and slowly presented in depth to a user depending on one's learning pace and expertise which also be a wonderful personal experience.

References

[1] R. Paulus et al. *A Deep Reinforced Model for Abstractive Summarization*. May 2017.

[2] R. Nallapati et al. *Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond*. [*Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL)*]. Aug 2016.

[3] A. See et al.
Get To The Point: Summarization with Pointer-Generator Networks Apr 2017

[4] A. M. Rush et al. *A Neural Attention Model for Abstractive Sentence Summarization*. Sep 2015.

[5] A. Vaswani et al., *Attention Is All You Need*. Jun 2017

[6] Rafal J. et al. *Exploring the Limits of Language Modeling*. Feb 2016.

[7] O. Vinyals et al. *Pointer Networks*. Jan 2017.

[8] Z. Zhang et al. *Stack-based Multi-layer Attention for Transition-based Dependency Parsing*. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Sep 2017.

[9] R. Nallapati et al. *SummaRuNNer: A Recurrent Neural Network based Sequence Model for Extractive Summarization of Documents*. Nov 2016.

[10] F. Liu et al. *Toward Abstractive Summarization Using Semantic Representations*.

[11] Deep Learning for Chatbots
http://www.wildml.com/2016/04/deep-learning-for-chatbots-part-1-introduction/

List of Figures