since 1995

Data Science Natural Language Processing NLP Papers Summary

Day 144: NLP Papers Summary – Attend To Medical Ontologies: Content Selection For Clinical Abstractive Summarization

By Ryan 23rd May 2020 No Comments

Objective and Contribution

Traditional abstractive text summarisation has the main problem of selecting key information from the source document. This paper proposed a method to content selection for clinical abstractive summarisation by incorporating salient ontological terms into the summariser. Content selection is treated as a word-level sequence-tagging problem. This has proven to improve on the SOTA results based on the MIMIC-CXR and OpenI datasets. We als uired

experts evaluation and shown that our approach generated a good quality summa in comparison to ground-truth.

Summarisation of Radiology Reports

Radiology reports contain of two important sections: FINDINGS and IMPRESSION. FINDINGS consists of the detailed observations and interpretation of the imaging study whereas IMPRESSION summarises the most critical findings. In the industry, most clinicians only read the IMPRESSION section as they have limited time to review the lengthy FINDINGS section. The automation and improvement of generation of IMPRESSION could significantly improve the workflow of radiologists.

Methodology

Our proposed model has two main components:

1. Content selector
2. Summarisation model

CONTENT SELECTOR

This component aims to select the most important ontological concepts within the report, specifically the FINDINGS section. This can be treated as word-level extraction task where we would like to extract words that are likely to be included in the IMPRESSION section. In practical, each word is tag with 1 if it meets two criteria:

1. The word is an ontology term
2. The word was directly copied into IMPRESSION

This allows us to capture the copy likelihood of each word and we used this to measure the importance of the word. The overall architecture is a biLSTM on top of a BERT embeddings layer (to take advantage of contextualised embeddings) and during inference time, our content selector will output the selection probability of each token in our source sequence.

SUMMARISATION MODEL

Our summarisation model has two encoders and a decoder (see figure below):

1. *Findings Encoder*. This is a biLSTM that takes in the word embeddings in FINDINGS section and generates a encoded hidden representation
2. *Ontology Encoder*. This is a LSTM that takes in identified ontology terms (by our content selector) and generates a fix context vector, our ontology vector
3. *Impression Decoder*. This is a LSTM that generates the IMPRESSION

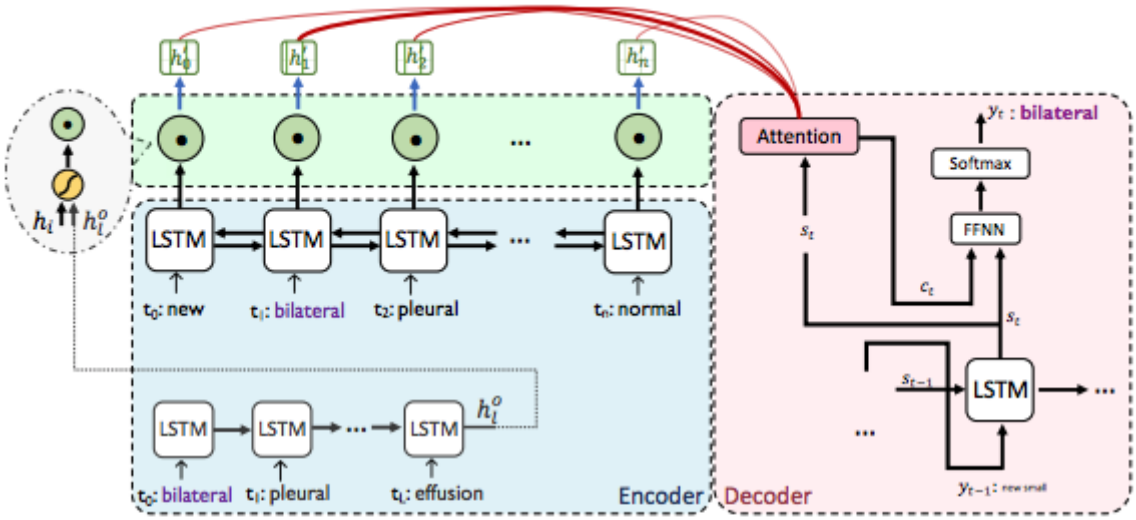

Figure 1: Overview of our summarization model. As shown, "bilateral" in the FINDINGS is a significant ontological term which has been encoded into the ontology vector. After refining FINDINGS word representation, the decoder computes attention weight (highest on "bilateral") and generates it in the IMPRESSION.

Next, we have a filtering gate that refine FINDINGS word representations using the ontology vector to produce ontology-aware word representations. The filtering gate concatenates, at each step, the current hidden state of word x and the fix ontology vector and process these through a linear with sigmoid activation function. To compute the ontology-aware word representations, we then take the output of the filtering gate and perform element-wise multiplication with the current hidden state of word x.

Our decoder is an LSTM that generates the IMPRESSION. The decoder will compute the current decoding state using the previous hidden state and previous generated tokens. The d r will also use the current decoding state to compute the attention distribution over the ontology-

aware word representations. The attention distribution is then used to compute the context vector. Finally, the context vector and the current decoding state is feed into a feed-forward neural network to either generate the next token or copy from FINDINGS.

Experiments and Results

We have two evaluation datasets: MIMIC-CXR and OpenI. MIMIC-CXR has 107372 radiology reports and OpenI has 3366 reports. For the radiology lexicon, we used RadLex, which consists of 68534 radiological terms.

MODELS COMPARISON

We have two extractive summarisation models (LSA and NEUSUM) and three abstractive summarisation models (Pointer-Generator (PG), Ontology-aware PG, and BOTTOMSUM). The BOTTOMSUM is the most relevant to our architecture as it utilises a separate content selector for abstractive text summarisation.

RESULTS

As shown in the table 1 above, our model significantly outperformed all the extractive and abstractive baseline models. Abstractive models significantly outperformed the extractive one indicating that human-written summary are formed abstractively and not just selecting sentences from the source. The difference in ROUGE performance between PG and Ontology-aware PG showcase the effectiveness and usefulness of incorporating salient ontological terms in summarisation model. As expected, BOTTOMSUM achieve the best results among the baseline models as it has the most similar architecture as our model. We believe the reason our model outperformed BOTTOMSUM is because we have an intermediate stage of refining word

representation based on ontological word. The table 3 below showcase the bene of incorporating content selection to summarisation model.

To evaluate the generalisation of our model, we also evaluate our model on OpenI against BOTTOMSUM and the results is showcase below in table 2. As shown, our model is also able to outperformed BOTTOMSUM in OpenI, illustrating the generalisation of our model.

EXPERT EVALUATION

Here, we randomly sampled 100 generated IMPRESSIONs with their associated gold summaries. We asked three experts to score the IMPRESSIONs on a scale of 1 – 3 (3 being the best) on readability, accuracy, and completeness. The results are display in the figure below. We observed that there are over 80% generated IMPRESSIONS that are scored as good as the associated human-written IMPRESSIONS. 73% and 71% of our IMPRESSIONS scored 3 on readability and accuracy and ties with human-written IMPRESSIONS, however only 62% of our IMPRESSIONS scored 3 on completeness. We believe this is due to the subjectiveness of what it's deem to be important in findings. Overall, it seems that our generated IMPRESSIONS are of high quality, however, there are still a gap between generated IMPRESSIONS and human-written ones.

Source: https://arxiv.org/pdf/2005.00163.pdf

Ryan

Data Scientist

Previous Post

Day 143: NLP Papers Summary - Unsupervised Pseudo-Labeling for Extractive Summarization on Electronic Health Records

Next Post

Day 145: NLP Papers Summary - SUPERT: Towards New Frontiers in Unsupervised Evaluation Metrics for Multi-Document Summarization

Data Science Natural Language Processing NLP Papers Summary

Day 365: NLP Papers Summary – A Survey on Knowledge Graph Embedding

Ryan
30th December 2020

Data Science Implementation Natural Language Processing

Day 364: Ryan's PhD Journey – OpenKE-PyTorch Library Analysis + code snippets for 11 KE models

Ryan
29th December 2020

Data Science Ryan's PhD Journey

Day 363: Ryan's PhD Journey – Literature Review – Knowledge Acquisition – 1st Passes XIV

Ryan
28th December 2020