Summaries of some papers by LCS2

Jeevesh Juneja

December 8, 2020

1 Neural Abstractive Summarization with Structural Attention

This paper [CKC20] tries to extend the structural attention of [Kim+17] and [Yan+16] to the task of summarising contextual question answering(CQA) data present on various websites like Quora, Yahoo! Answers etc. They hypothesize that these methods would prove useful in modelling long-term structural dependencies present in the CQA data-sets, and obtain results that support the same.

1.1 Structural Attention

We know that the attention weights of [BCB16], can be seen as a latent variables that need to be jointly learnt along with sequence to sequence learning task under consideration. The contextual attention used in [BCB16] takes a linear combination of the token-wise vector representations(obtained via some model like LSTM) to obtain a contextual representation for each token. This linear combination is a weighted sum:

- where the coefficients must be positive and sum to 1(so that they can be seen as a probability distribution), and
- each coefficient is predicted by a NN which takes as input the vector representation of the words, between which attention is to be determined. ($O(N^2)$[1] prediction corresponding to each pair of words)

In the following "attention weights" can refer to un-normalized numbers predicted by the NN model, or their exponentiated version or normalized version, whichever is appropriate.

The second point above, is not much of a constraint, but the first one is a real constraint on the space of latent variables(attention weights) that can be used with the model. With this constraint, and the use of soft-max, the end result is that there is an inductive bias in the model to encourage it to chose those elements in the latent variable space, that clearly attend to a lesser number of positions in the vector representation of the input.

[Kim+17] explore further in what ways could we constraint the latent variable space of the attention mechanism. Their construction is as follows : Instead of passing $O(N^2)$ numbers predicted by the NN, through a soft-max and considering them as N probability distributions,

- they assume the existence of a CRF upon the various elements of latent space, i.e., space of attention weights[2]
- they consider the predicted numbers as the log potentials of various connections in the CRF,
- Additionally, this CRF restricts the attention weights to be non-zero only for a set of pairs of words that can form a projective dependency tree.
- then, they calculate the margin probability (using Inside-Outside algorithm) of each element of the latent space under the CRF and use that probability as the final attention weight.

[1] where N is the number of tokens in the sentence

[2] the space of vectors of size NXN, where v_{ij} is the attention given to the j^{th} word when making the context vector of the i^{th} word

In effect, their model has an inductive bias to restrict the information flow, to happen only along the edges of a projective dependency tree over the input sequence. This is a stronger inductive bias than in the simple contextual attention described before. But, since this restriction is a soft one, rather than a hard one, the multi-valued function represented by structural attention form a super-set of those represented by contextual attention.
The above approach, though, is more restrictive than we'd like it to be. We may want the information to flow through a non-projective tree, to allow more complex dependencies. [LL17] try to do the same. The basic problem is that the recursion of inside-outside algorithm is only applicable to trees generated from a context free grammar, and non-projective trees can't be generated using CFGs. The basic problem is to find the partition function of the CRF, that is the sum of potentials of all the directed spanning trees(projective as well as non-projective). The potential of a DST is defined as the product of potentials of all the edges in the DST. Their solution is equivalent to the the Dynamic Programming formulation of the following recursion :

$$T(m, G_n) = \sum_{v \in G_{n-1}} edge(m, v) T(v, G_{n-1}) \tag{1}$$

where $T(m, G_n)$ is the sum of potentials of directed spanning trees rooted at node m in the original graph of n nodes; the summation is over all nodes in the graph of $n-1$ vertices(G_{n-1}) obtained by deleting node m from G_n; $edge(m, v)$ is the potential of the directed edge from node m to node v.

This can also be stated in terms of the a modified Laplacian matrix of the given graph and derived as a corollary of Matrix Tree Theorem, as in [Koo+07].

[CKC20] directly uses this kind of structural attention, to help their pointer generator model to capture better structural representations of the input, with better contexts. Their experiments show that including SA, an improvement in ROGUE-1 score can be obtained over simple pointer generator network. Moreover, the gap widens as we try to use the models on longer and longer documents.

1.2 Hierarchical Modelling and Copy Mechanism

[CKC20], in order to capture even longer term dependencies, makes the PG network hierarchical, inspired from [Yan+16]. This approach is also more helpful in integrating varied information from multiple documents(leads to better improvement on CQASUMM than on MultiNews). It basically consists of pooling together representations of all the words in a document(an answer) to obtain answer-level representations and then performing SA on them to obtain structured representations of those answer-level representations. Then these are used to generate a summary of all the documents. The paper also includes the copy mechanisms of [SZL18] which brings about a rather impressive jump in the ROGUE scores.

1.3 Possible improvements

Regarding the CQA task, there is a possibility of the information in the question with each of the documents. Also, there seems to be a lot of possible directions / experiments still remaining regarding structural attention. For e.g., we can restrict the attention weight matrix(NXN) to actually be sparse, or use Gumbel soft-max instead of simple soft-max; this would lead to even better interpret-ability of the model. Also we can try multi-headed architecture, as opposed to current single headed one. Another possible direction relates to increasing the depth up-to which the information is allowed to flow in the directed spanning trees of SA(currently they allow for at most one level up, i.e., collecting information from parents and one level down, collecting information from children). Regarding hierarchical SA, it seems natural to try to explore variants that use combined form of attention at various encoder levels in various decoder levels. Also, one can try to refine the attention choices at lower levels of hierarchy, by using the information available at layers higher in the hierarchy.This can be done by constructing an architecture composed of alternating fine-grained and broader-context layers. Also, hierarchical structured attention modules can be pre-trained by making the higher level layers generate back the contents of the lower layer. Token level SA can be pre-trained using the various tree-bank data available.

2 Deep Exogenous and Endogenous Influence Combination for Social Chatter Intensity Prediction

The paper [Dut+20], is a beautiful attempt at trying to predict user engagement on a post independent of the network structure, based only on the content of endogenous(within the forum) and exogenous(like news etc. which are expected to influence the chatter within the forum) posts.

2.1 The Model

The model they present, is really well-designed and optimized for their setting. Some really good choices are as follows :

- Using only simple convolutions(of kernel size 1,3,5) for forming representation of content of a post/news article, since the task doesn't require much detailed information, and predictions can be made based only on broader features of the document.
- Using GRU to aggregate the representations of posts/news articles rather than LSTM.
- Maintaining separate GRU for posts and news content.
- Using Time Evolving Convolutions rather than recurrent units; as posts in a small interval may not be directly related to one another, but depend on the aggregated representations of news/posts of previous intervals.
- The mechanism corresponding to observing comments on a post for a short time and predicting the following chatter is also, integrated beautifully into the whole pipeline, using an LSTM that takes as inputs the number of comments in each of the small observation windows till now.

In addition to the above things, they also include the information about which sub-reddit is under consideration, via a scaling factor(for the output of their model) that captures the average commenting activity on that sub-reddit.

All this, brings home the fact that small well-calibrated and well-designed models can solve problems more efficiently than bigger ones, if the task is supports it.

2.2 The Experiments

The authors carry out an extensive set of ablation experiments that clearly show that each of the components(endogenous as well as exogenous) of their model leads to significant increase in performance. Also, similar to previous models, their model degrades when predicting bigger cascades, but sort of recovers if given large enough observation window. They also show how the importance of exogenous signal, endogenous signal and early observation changes with the sub-reddit under consideration. So,they probably should have used the sub-reddit tag s_j^v to form a combination of G_k^n and G_k^s rather than simple concatenation.

2.3 Possible Improvements

Many times, in models that try to predict the aggregate(over some prediction window) of some quantity, there tends to be a systematic bias that under-estimate or over-estimate the actual values, this effect aggravates as the prediction window increases. The MAPE error makes this effect invisible, but can be successfully seen and corrected via Mean Bias Error(MBE). The authors could have included that probably.

Another obvious direction is to improve the performance of the model for longer prediction windows. For that, I think it is critical to strengthen the comment aggregation model, while allowing it to capture the the endogenous and exogenous signals directly as well as the inherent tree structure of comments on Reddit.

References

[BCB16] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. *Neural Machine Translation by Jointly Learning to Align and Translate*. 2016. arXiv: `1409.0473 [cs.CL]`.

[CKC20] Tanya Chowdhury, Sachin Kumar, and Tanmoy Chakraborty. *Neural Abstractive Summarization with Structural Attention*. 2020. arXiv: `2004.09739 [cs.CL]`.

[Dut+20] Subhabrata Dutta et al. *Deep Exogenous and Endogenous Influence Combination for Social Chatter Intensity Prediction*. 2020. arXiv: `2006.07812 [cs.SI]`.

[Kim+17] Yoon Kim et al. “Structured Attention Networks”. In: *CoRR* abs/1702.00887 (2017). arXiv: `1702.00887`. URL: `http://arxiv.org/abs/1702.00887`.

[Koo+07] Terry Koo et al. “Structured Prediction Models via the Matrix-Tree Theorem”. In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 141–150. URL: `https://www.aclweb.org/anthology/D07-1015`.

[LL17] Yang Liu and Mirella Lapata. “Learning Structured Text Representations”. In: *CoRR* abs/1705.09207 (2017). arXiv: `1705.09207`. URL: `http://arxiv.org/abs/1705.09207`.

[SZL18] Kaiqiang Song, Lin Zhao, and Fei Liu. “Structure-Infused Copy Mechanisms for Abstractive Summarization”. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 1717–1729. URL: `https://www.aclweb.org/anthology/C18-1146`.

[Yan+16] Zichao Yang et al. “Hierarchical Attention Networks for Document Classification”. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, June 2016, pp. 1480–1489. DOI: `10.18653/v1/N16-1174`. URL: `https://www.aclweb.org/anthology/N16-1174`.