Mortality Prediction Models with Clinical Notes Using Sparse Attention at the Word and Sentence Levels

Miguel Rios and **Ameen Abu-Hanna**
Department of Medical Informatics, Amsterdam UMC,
University of Amsterdam, The Netherlands
{m.a.riosgaona,a.abu-hanna}@amsterdamumc.nl

Abstract

Intensive Care in-hospital mortality prediction has various clinical applications. Neural prediction models, especially when capitalising on clinical notes, have been put forward as improvement on currently existing models. However, to be acceptable these models should be performant and transparent. This work studies different attention mechanisms for clinical neural prediction models in terms of their discrimination and calibration. Specifically, we investigate sparse attention as an alternative to dense attention weights in the task of in-hospital mortality prediction from clinical notes. We evaluate the attention mechanisms based on: i) local self-attention over words in a sentence, and ii) global self-attention with a transformer architecture across sentences. We demonstrate that the sparse mechanism approach outperforms the dense one for the local self-attention in terms of predictive performance with a publicly available dataset, and puts higher attention to prespecified relevant directive words. The performance at the sentence level, however, deteriorates as sentences including the influential directive words tend to be dropped all together.

1 Introduction

Deep learning has become a promising approach for clinical prediction models (Rajkomar et al., 2018; Shickel et al., 2019). Moreover, natural language processing (NLP) applications based on neural network (NN) models have shown the potential to benefit the task of mortality prediction (Kemp et al., 2019). However, NN models based on clinical notes are difficult to interpret, due to their black-box nature. Decision makers need transparency in the way words in an input clinical note contribute to the overall prediction. One approach to gain transparency is the attention mechanism, which have shown improvements across NLP applications (Vaswani et al., 2017; Devlin et al., 2019) for text representation. The attention mechanism assigns weights to an input vector representation of a layer in a NN (Bahdanau et al., 2015), and it has been used to identify the relative importance of a word (Clark et al., 2019; Huang et al., 2019; Vashishth et al., 2019).

Caicedo-Torres and Gutiérrez (2020) propose a mortality prediction model with clinical notes based on a convolutional architecture, where the max pooling layer is used to inform word importance. Lovelace et al. (2019) incorporate an attention mechanism on top of the convolutional representation model to highlight words from the input clinical notes.

Current attention mechanisms produce a dense weight distribution over the input vector representation. The dense attention will assign a weight to all words in the context even if they are irrelevant for the current prediction. Moreover, attention weights may not be interpretable given that their alterations do not lead to a change in predictions (Serrano and Smith, 2019; Jain and Wallace, 2019). In contrast, sparse attention can lead to more accurate models and more transparent presentation by assigning zero weights to some words (Martins and Astudillo, 2016; Niculae and Blondel, 2017; Correia et al., 2019).

In this paper, we investigate the predictive performance of mortality prediction models, in terms of discrimination and calibration, using clinical notes. In particular, we study the effect of sparse versus dense attention mechanisms within a hierarchical architecture: self-attention is applied at the local level to weigh words in a sentence, and at the global level to weigh sentences. We show preliminary results for the task of in-hospital mortality prediction

using a publicly available dataset. First, we compare the predictive performance between a model based on local self-attention over words using the standard dense attention, and two sparse attention mechanisms sparsemax and entmax. The sparse attention mechanisms show better performance compared to dense attention. A preliminary evaluation showed that sparsemax consistently identified prespecified directive words known to be associated with mortality. Finally, we use the different attention mechanisms on a global self-attention model across sentences on clinical notes. In contrast to the previous experiment, the global attention model with the dense mechanism outperforms the sparse ones. The latter seems to throw sentences out with the directive words.

2 Sparse Attention

The Self-attention or scaled dot-product attention is a main component in current state-of-the-art NLP models (Vaswani et al., 2017). Self-attention computes the representation of a sequence by processing each of its positions and associating the position with other positions for indications of significance. The self-attention layer is defined as follows:

$$\text{Att}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \boldsymbol{\pi}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d}}\right)\boldsymbol{V} \qquad (1a)$$

where $\boldsymbol{Q}$ indicates queries, $\boldsymbol{K}$ keys, $\boldsymbol{V}$ values, and d vector dimensionality. The queries are a linear transformation of the input vector x with a parameter matrix W_q defined as $\boldsymbol{Q} = W_q x$, and similarly for keys $\boldsymbol{K} = W_k x$, and values $\boldsymbol{V} = W_v x$. The standard $\boldsymbol{\pi}$ mapping is based on softmax, and it is used to normalise the attention weights. Moreover, self-attention can be extended to multiple heads that potentially allows the model to focus on different positions (Vaswani et al., 2017).

Martins and Astudillo (2016) propose sparsemax as a differentiable mapping that provides a sparse alternative to softmax. Sparsemax projects the input vector into the probability simplex, where it is likely to be in the boundary of the simplex and then becoming sparse. Moreover, the entmax mapping produces sparse distributions by defining interpolations between softmax and sparsemax (Peters et al., 2019; Correia et al., 2019). The entmax becomes an intermediate mapping between sparsemax and softmax. Correia et al. (2019) propose a sparse attention mechanism for transformer architectures by replacing the π mapping by sparsemax, and entmax.

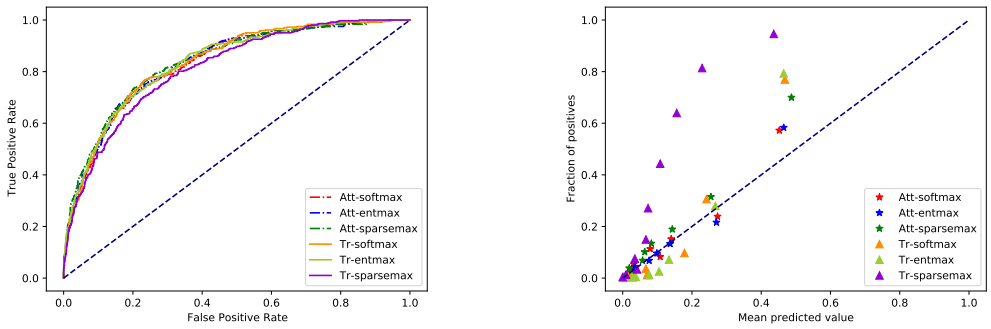
Most mortality prediction models based on free text represent multiple notes of an ICU stay as a sequence of sentences. The model hierarchically composes word representations (i.e. word encoder) into a sentence, and sentences (i.e. sentence encoder) into a patient representation for predicting mortality (Grnarova et al., 2016; Si and Roberts, 2019). For example, Grnarova et al. (2016) use a hierarchy of convolutional neural networks (CNN) on the word and sentence encoders for mortality prediction given complete ICU stays.

We use a word encoder with one layer self-attention and one head as a baseline to represent local dependencies over words in a sentence. The clinical notes of a patient for an ICU stay $x_{i,t}$ consist of i words and t sentences. The input $x_{i,t}$ is first represented with pre-trained word embeddings, and then projected into a linear layer followed by self-attention (Eq. 1) to encode words of each sentence. Finally, the features of the word encoder are summarised by averaging to produce a patient representation for a linear prediction layer. For the local self-attention, we define the following models: **Att-softmax** as baseline with softmax mapping for dense attention weights, **Att-sparsemax** with the attention weights based on the sparsemax mapping, and **Att-entmax** with the entmax mapping.

A limitation of a local architecture with self-attention is the lack of long distance dependencies across sentence representations. We use a hierarchical architecture for modelling long distance relations within an ICU stay. We replace the CNN with a transformer layer (Vaswani et al., 2017), which is based on self-attention and positional embeddings, on both levels of the hierarchy (Pappagari et al., 2019; Zhang et al., 2019). For the global attention, we define the following hierarchical transformers: **Tr-softmax**, **Tr-entmax**, and **Tr-sparsemax**. We describe the self-attention and hierarchical transformer architectures in more detail in Appendix A.

Model	AUC-ROC↑	AUC-PR↑	Brier↓
Att-softmax	0.824 ± 0.003	0.435 ± 0.008	0.085 ± 0.001
Att-entmax	0.827 ± 0.004	0.445 ± 0.010	0.084 ± 0.002
Att-sparsemax	0.834 ± 0.004	0.467 ± 0.011	0.087 ± 0.001
Tr-softmax	0.839 ± 0.005	0.462 ± 0.005	0.093 ± 0.016
Tr-entmax	0.828 ± 0.007	0.433 ± 0.014	0.114 ± 0.004
Tr-sparsemax	0.801 ± 0.009	0.373 ± 0.025	0.196 ± 0.076

Table 1: In-hospital mortality results on the test dataset with 5 runs ± standard deviation on each model reporting AUC-ROC, AUC-PR, and Brier score.



(a) (b)

Figure 1: Receiver operating characteristic curve (a) and calibration curve (b) for in-hospital mortality.

3 Experiments

The Medical Information Mart for Intensive Care (MIMIC-III) database includes ICU information such as demographics, vital measurements, laboratory test results, procedures, medications, caregiver notes, imaging reports, and mortality for critical care patients (Johnson et al., 2016). Harutyunyan et al. (2019) propose a public benchmark from MIMIC-III for modelling mortality, length of stay, physiologic decline, and phenotype classification. We use the in-hospital mortality benchmark cohort for extracting clinical notes in English based on the first 48 hours of an ICU stay to develop a prediction model. The cohort excludes ICU stays with missing length-of-stay, patients under 18, multiple ICU stays, stays under 48 hours, and without observations on the first 48 hours. The mortality class is defined by comparing the date of death across hospital admissions and discharge times with a resulting mortality rate of 13.2%. To preprocess the extracted notes, we tokenize with NLTK[1], lowercase, and exclude duplicate and discharge notes. The training/validation/test datasets consist of data from 14,681, 3,222 and 3,236 patients, respectively. We report the mean and standard deviation with 5 random runs for the area under the receiver operator characteristic curve (AUC-ROC), area under the precision-recall curve (AUC-PR), and the Brier score.

We pretrain GloVe (Pennington et al., 2014) word embeddings on the benchmark training split with: 100D, minimum frequency of 5 words, window of 15 words, 20 epochs, and further fine-tune them on the downstream task. For the prediction model, we use the following hyperparameters: Adam optimiser (Kingma and Ba, 2014), learning rate 1e−4, epochs 30, hidden size 128, batch size 16, and dropout 0.2. For each training instance the max number of words i in a sentence is 50 and the max number of sentences t is 1000. Finally, we perform hyper-parameter and model selection tuning with the validation dataset based on the AUC-ROC.

3.1 In-hospital Mortality Results

We report the performance on the task of in-hospital mortality prediction using sparse attention, and compare it to dense attention. Table

[1]NLTK tokenizer for English: https://www.nltk.org/index.html

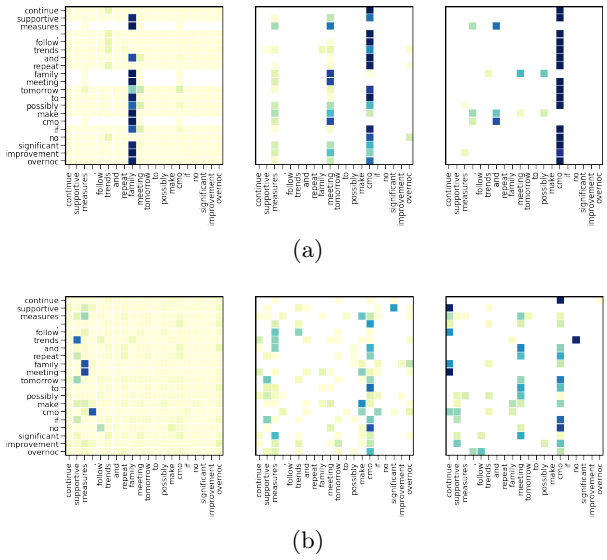


Figure 2: Attention heatmaps corresponding positive class instance. Att-softmax, Att-entmax and Att-sparsemax respectively in (a). Tr-softmax, Tr-entmax and Tr-sparsemax in (b).

1 shows the AUC-ROC, AUC-PR, and Brier score results of the validation and test datasets. The Att-sparsemax shows competitive results compared to entmax and softmax on both metrics AUC-ROC and AUC-PR. The Brier score increases for the sparse attention models, where the models with sparse attention assign higher output probabilities to both classes. Figure 1 shows the ROC and calibration curves. The sparse attention models show competitive performance compared to the baseline on ROC. However, we observe from the calibration curve that the sparse attention underestimates predictions for the local attention models.

Moreover, the addition of sparse attention on the second level of the hierarchy for the transformer models, hurts the performance measures and calibration. The sparse transformer models show a marked miscalibration in Figure 1 where it underpredicts the true probabilities.

3.2 Qualitative Analysis

We show example attention heatmaps from the local self-attention model for sentences that include any of the following directives: do not resuscitate (dnr), do not intubate (dni), and comfort measures only (cmo). The directive words are known to be strongly correlated with mortality. For example, if the family of a patient has decided for comfort measures only then this refers to palliative treatment of a patient thought to be dying.

Jain and Wallace (2019) define explainable attention when the inputs with the highest attention weights are relevant for a given prediction. Figure 2 shows attention heatmaps for two sentences that contain any of the directive words. The heatmaps denote the relative importance for each word embedding (row) with respect to its context (column), the darker the cell, the higher the weight. One can observe that the sparse attention put most of its weights on the directive words, and that the highest attended words differ between the sparse and dense mechanisms. However, in Figure 2 (a) entmax better distributes its weights for cmo. We show more attention examples in Appendix B. On the hierarchical transformer the sentence level sparsemax is consistently too aggressive in selecting weights. For example, in the sentence from Fig. 2 (b) the model only assigns attention to 8 out of 174 sentences. Moreover, the attention weights at the word encoder level shifted away from the directive words compared to the local attention.

4 Discussion and Future Work

In this paper, we compared the standard dense attention with sparse attention mechanisms on two setups of local and global dependencies across representations for mortality prediction with clinical notes. The sparsemax attention mechanism shows competitive predictive performance compared to the dense attention weights on the local self-attention model. However, the global model based on sparse attention produces under-confident predictions. The drop in performance in the hierarchical transformer is due to the sparse attention on the sentence encoder only assigning weights to few sentences, where most ICU stays contain hundreds of sentences.

Future work meriting investigation includes combining sparse and dense mechanisms on the hierarchical transformer. In addition to showing examples of the attention heatmaps, one may also analyse the predicted weights by measuring their correlation to feature importance metrics (Vashishth et al., 2019; Serrano and Smith, 2019).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

William Caicedo-Torres and Jairo A. Gutiérrez. 2020. Iseeu2: Visually interpretable ICU mortality prediction using deep learning and free-text medical notes. *CoRR*, abs/2005.09284.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Gonçalo M. Correia, Vlad Niculae, and André F. T. Martins. 2019. Adaptively sparse transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2174–2184, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Paulina Grnarova, Florian Schmidt, Stephanie L. Hyland, and Carsten Eickhoff. 2016. Neural document embeddings for intensive care patient mortality prediction. *CoRR*, abs/1612.00467.

Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. 2019. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1).

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv:1904.05342*.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035.

J. Kemp, Alvin Rajkomar, and Andrew M. Dai. 2019. Improved patient classification with language model pretraining over clinical notes. *ArXiv*, abs/1909.03039.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. Cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

Justin R. Lovelace, Nathan C. Hurley, Adrian D. Haimovich, and Bobak J. Mortazavi. 2019. Explainable prediction of adverse outcomes using clinical notes. *CoRR*, abs/1910.14095.

Andre Martins and Ramon Astudillo. 2016. From softmax to sparsemax: A sparse model of attention and multi-label classification. volume 48 of *Proceedings of Machine Learning Research*, pages 1614–1623, New York, New York, USA. PMLR.

Vlad Niculae and Mathieu Blondel. 2017. A regularized framework for sparse and structured neural attention. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Raghavendra Pappagari, Piotr Zelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. 2019. Hierarchical transformers for long document classification. *CoRR*, abs/1910.10781.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.

Ben Peters, Vlad Niculae, and André F. T. Martins. 2019. Sparse sequence-to-sequence models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1504–1519, Florence, Italy. Association for Computational Linguistics.

Alvin Rishi Rajkomar, Eyal Oren, Kai Chen, Andrew Dai, Nissan Hajaj, Mila Hardt, Peter J. Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, Patrik Per Sundberg, Hector Yee, Kun Zhang, Yi Zhang, Gerardo Flores, Gavin Duggan, Jamie Irvine, Quoc Le, Kurt Litsch, Alex Mossin, Justin Jesada Tansuwan, De Wang, James Wexler, Jimbo Wilson, Dana Ludwig, Samuel Volchenboum, Kat Chou, Michael Pearson, Srinivasan Madabushi, Nigam Shah, Atul

Butte, Michael Howell, Claire Cui, Greg Corrado, and Jeff Dean. 2018. Scalable and accurate deep learning for electronic health records. *npj Digital Medicine*.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Benjamin Shickel, T. Loftus, Lasith Adhikari, T. Ozrazgat-Baslanti, A. Bihorac, and Parisa Rashidi. 2019. Deepsofa: A continuous acuity score for critically ill patients using clinically interpretable deep learning. *Scientific Reports*, 9.

Yuqi Si and Kirk Roberts. 2019. Deep patient representation of clinical notes via multi-task learning for mortality prediction. *AMIA Summits on Translational Science Proceedings*, 2019:779.

Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. 2019. Attention interpretability across NLP tasks. *CoRR*, abs/1909.11218.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

Xingxing Zhang, Furu Wei, and Ming Zhou. 2019. HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5059–5069, Florence, Italy. Association for Computational Linguistics.

A Architectures

We define the following architectures given an input $x_{i,t}$ that consist of clinical notes for an ICU stay represented into i words and t sentences.

Self-attention Figure 3 shows the architecture for the self-attention model.

Hierarchical Transformer Figure 4 shows the architecture for the hierarchical transformer model. The architecture consists of pre-trained emebddings, position embeddings for words $\omega_{\text{pos\_i}}$, sentences $\omega_{\text{pos\_t}}$, and transformer layers with self-attention.

B Attention Heatmaps

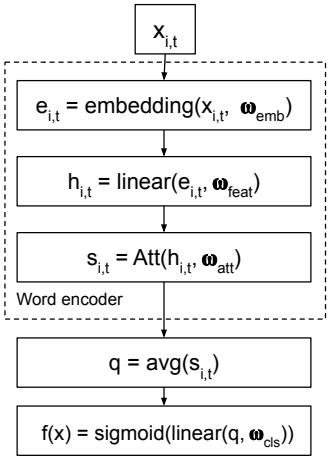


Figure 3: Self-attention architecture.

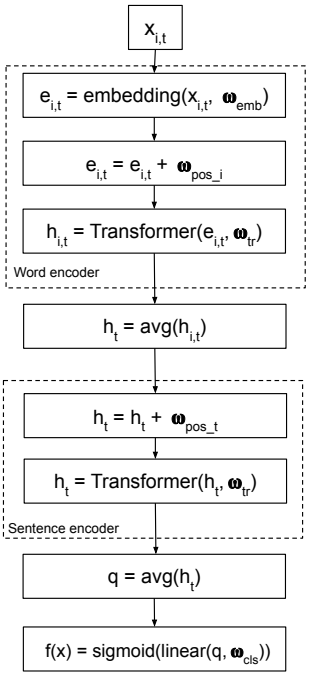
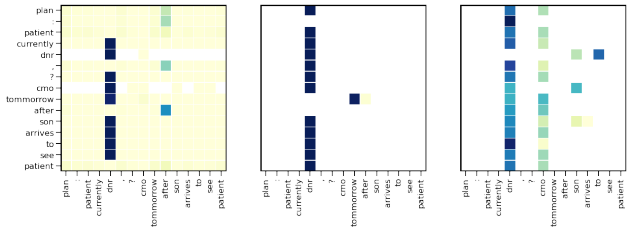
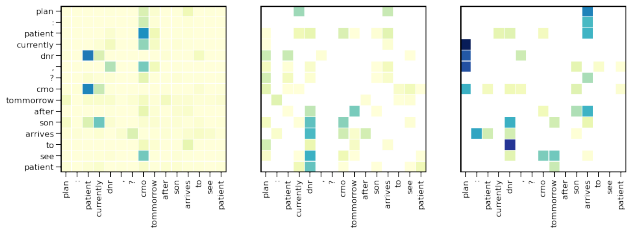


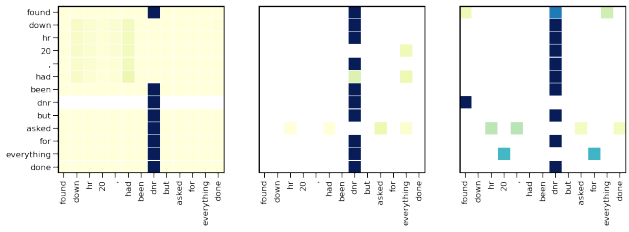
Figure 4: Hierarchical transformer architecture.



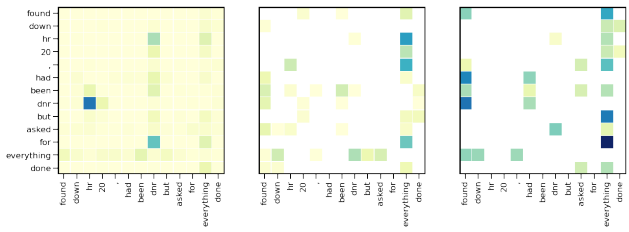
(a)



(b)



(c)



(d)

Figure 5: Attention heatmaps for sentences in a positive instance (a) Att-softmax, Att-entmax and Att-sparsemax respectively, and (b) Tr-softmax, Tr-entmax and Tr-sparsemax. Sentences in negative instance (c) Att-softmax, Att-entmax and Att-sparsemax and (d) Tr-softmax, Tr-entmax and Tr-sparsemax .