Identifying Diagnostic Test Accuracy Publications using a Deep Model.

Gaurav Singh[1], Iain Marshall[2], James Thomas[1], and Byron Wallace[3]

[1] UCL, UK
gaurav.singh.15@ucl.ac.uk
[2] Kings College London, UK
[3] Northeastern University, USA

1 Abstract

In this work, we used a deep model architecture to identify DTA studies pertaining to a given review topic. We were provided the list of relevant documents selected based on abstracts and full text for different reviews topics. We extracted the abstract and title to be used as features to describe those documents, and learned the deep neural net model that takes as input the abstract and title of the studies, and topic of the review to obtain a binary classification of whether that study is a relevant DTA to the review in question.

2 Model

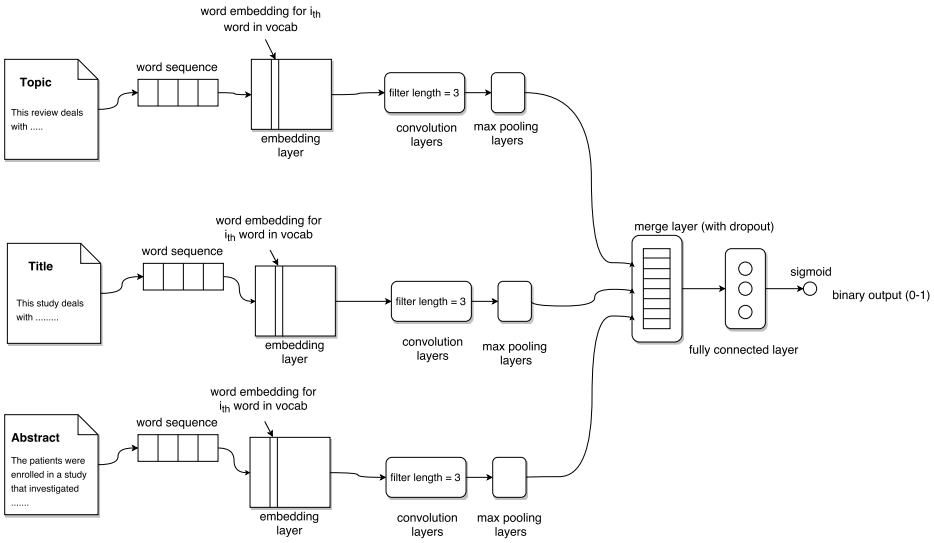

Fig. 1: Deep Model Architecture used for the Task.

The proposed model takes as input the title and abstract of the paper as sequences of words. These are then fed into the embeddings layer that outputs a matrix of words vectors corresponding to the given words. It is then passed through a 1-dimensional convolution layer of filter length 3. Similarly, the topic of the review in question is also passed through the embedding layer and into the convolution layer of filter length 3. The embeddings generated by the three different convolution layers are then merged, and passed through a dense fully connected layer with dropout, and sigmoid activation function for output. The loss function used at the output layer is binary cross-entropy.

2.1 Tuning

All the parameters were tuned on a held out validation dataset. The probabilities of dropout were tuned over a range of 10 equidistant values in the interval $[0, 1]$. The optimal value of dropout probability obtained was 0.6. The structure of the network was also trained on the held out validation dataset. We experimented with different filter lengths, and different number of convolution layers.

3 Results

We can see the performance of the model on the held out dataset in Figure 2. We can observe that the model managed to work much better than a random classifier would have performed. We can see the macro-averaged performance of the model in identifying relevant abstracts, and relevant full text documents in Table 1. We can see the micro-averaged performance of the model in identifying relevant abstracts and relevant full text documents in Table 2, obtained using the script provided for evaluation.

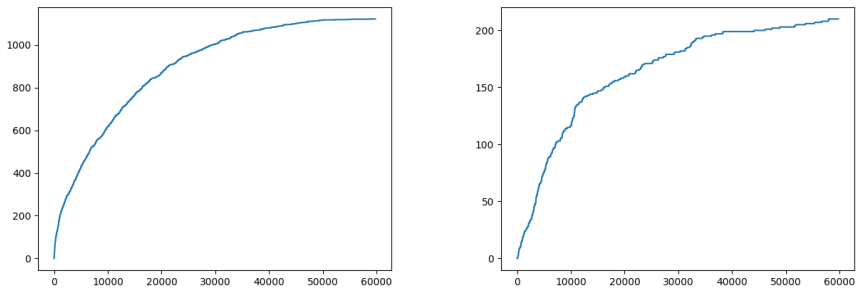

Fig. 2: It plots the number of relevant documents identified based on abstracts versus the number of documents manually annotated (left), and the number of relevant documents identified based on full text versus the number of documents manually annotated (right). It is based on the held out data during training.

Accuracy	0.79537	Accuracy	0.99482
AUC	0.56379	AUC	0.57593
WSS @ 95.0 %	0.08171	WSS @ 95.0 %	0.14705
WSS @ 100.0 %	0.00083	WSS @ 100.0 %	0.00335

Table 1: Results on the test set for identifying relevant abstracts (left), and results on the test set for identifying relevant studies (right). In both cases, we use the abstract and title of the paper, in addition to the review topic to identify the relevant studies. It is only different in the ground truth labels generated based on the abstract or the full text of the study. Note that these results are macro averages, and not micro averages across different reviews.

WSS@100	0.072	WSS@100	0.077
WSS@95	0.064	WSS@95	0.076
NCG@10	0.117	NCG@10	0.059
NCG@20	0.229	NCG@20	0.152
NCG@30	0.347	NCG@30	0.247
NCG@40	0.440	NCG@40	0.359
NCG@50	0.536	NCG@50	0.467
NCG@60	0.627	NCG@60	0.584
NCG@70	0.729	NCG@70	0.688
NCG@80	0.826	NCG@80	0.788
NCG@90	0.906	NCG@90	0.891
NCG@100	0.998	NCG@100	0.992
T. Cost	3918.733	T. Cost	3918.733
Norm Area	0.507	Norm Area	0.522

Table 2: Results on the test set for identifying relevant abstracts (left), and results on the test set for identifying relevant studies (right). In both cases, we use the abstract and title of the paper, in addition to the review topic to identify the relevant studies. It is only different in the ground truth labels generated based on the abstract or the full text of the study. Note that these results are micro averages over reviews obtained using the evaluation script provided.

4 Discussion

In previous work, we have built a classifier which, when presented with an unknown citation (i.e. title/abstract), can predict whether it describes a Randomized Controlled Trial (RCT) or not. Performance and technical details can be found in Wallace *et al.* [1]. The performance of this classifier on studies retrieved in searches for systematic reviews is good, and can reduce the manual screening burden by up to 80% while maintaining 100% recall. This is potentially very useful, but it is able to do this because: 1) it has been built on a large unbiased training dataset of 280,000 manually-labelled citations; and 2) the searches for systematic reviews of RCTs retrieve a large number of references which are not RCTs.

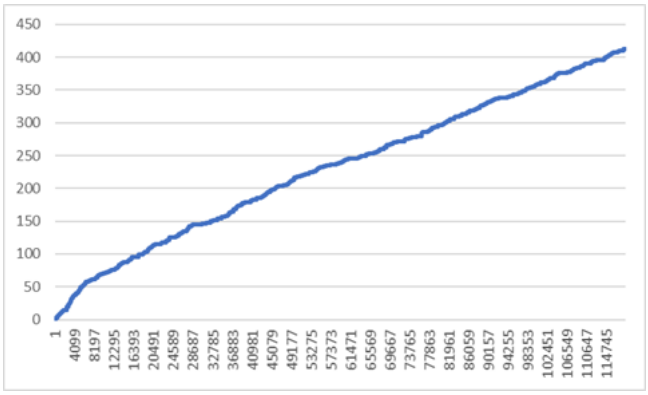

Fig. 3: It plots the number of relevant documents identified (as per full text) versus the number of abstracts manually annotated, using review independent DTA classifier.

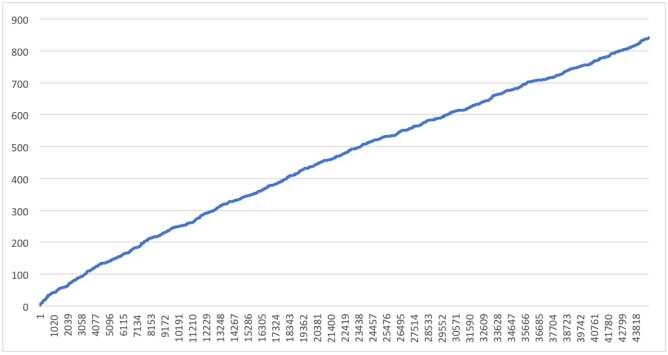

Fig. 4: It plots the number of relevant documents identified (as per abstract) versus the number of abstracts manually annotated, using review independent DTA classifier.

We appear to have a different situation with regards to DTA studies. We do not have the luxury of a large dataset on which to build a DTA classifier. The data presented for this exercise, for example, are the result of searches and screening decisions for DTA systematic reviews - rather than searches and screening decisions for DTAs. This means that the negative class in the DTA dataset contains large numbers of DTA studies, because they were irrelevant for the specific DTA review in question. This makes it impossible to use this dataset to build a generic DTA classifier. Moreover, we also built a DTA classifier from records we obtained outside this dataset - approximately 1,500 records which were manually labelled as to whether they described a DTA study or not. The results obtained, when using this classifier against the DTA training dataset for this task are shown in the Figure 3 and 4. Other than a small boost at

the bottom left of the graph in Figure 4, we can see that this classifier does not perform well. Especially, in comparison to the results of the deep model presented in the previous section.

5 Acknowledgements

JT and GS acknowledge support from Cochrane via the Transform project. BCWs contribution to the work was supported by the Agency for Healthcare Research Quality, grant R03-HS025024, and from the National Institutes of Health/National Cancer Institute, grant UH2-CA203711. IJM acknowledges support from the UK Medical Research Council, through its Skills Development Fellowship program, grant MR/N015185/1.

References

1. B. C. Wallace, A. Noel-Storr, I. J. Marshall, A. M. Cohen, N. R. Smalheiser, and J. Thomas. Identifying reports of randomized controlled trials (rcts) via a hybrid machine learning and crowdsourcing approach. *Journal of the American Medical Informatics Association*, 2017.