Explainable and Steerable learning for Medical Abstract Classification

Vikram Mandikal, Sundara Raman Ramachandran
UT EIDs: MMV894, SR47262

Abstract

Medical abstract classification involves tagging each sentence in the abstract as one of the following classes - *background, objective, method, result, conclusion*. In this work we develop explainable deep learning models for performing this task by incorporating prototypes in our architectures. These prototypes are trainable end-to-end like (Ming et al., 2019) and we observe that the learned prototypes correspond to typical sentences of these classes. We also use the steerable property of these prototypes to boost the performance of our model. Our model achieves a performance close to the state-of-the-art on the Pubmed RCT dataset 20k (Dernoncourt and Lee, 2017) while being explainable and steerable.

1 Introduction

From the year 1665, more than 50 million scholarly research articles have been published (Jinha, 2010) and each year approximately 2.5 million new scientific papers are getting published (Ware and Mabe, 2015). With this huge amount of literature corpus, extracting useful information efficiently has become very difficult due to its sheer amount. Thus, an automatic and intelligent tool to help users locate the information of interests quickly and comprehensively has become more than a necessity.

When searching for relevant literature in a certain field, investigators first skim through the abstracts of research papers to see whether they match the criterion of their interest. This process can be efficiently performed in a less time-consuming way if the abstracts are structured, i.e., if the structural elements of scientific abstracts such as purpose, methods, results, and conclusions (American National Standards Institute, 1979) are explicitly stated. In spite of the advantages of having structured abstracts, even today, a significant portion of the scientific abstracts are still unstructured, making information retrieval and comprehension a cumbersome process. In this work, we develop explainable deep learning models to address this challenge.

In a scientific abstract (especially medical), each sentence can be assigned to a structural element sequentially. Identifying the role of each sentence in the abstract can be formulated as a sequential sentence classification task, as the label of any particular sentence is associated with the labels of its neighboring sentences. This is in contrast to the general sentence classification problem, where each sentence is classified individually, and no contextual information can be used. Previous state-of-the-art method (Jin and Szolovits, 2018) implemented a bi-directional long short-term memory (bi-LSTM) layer over the representations of individual sentences so that it can encode the contextual content and semantics from preceding and succeeding sentences for better categorical inference of the current one.

Also, in recent years we have witnessed a surge in the use of deep neural networks for sequence modeling. But with the use of deep neural networks it is not very explicit & challenging to explain the rationales behind a particular model output. Moreover for sensitive data such as medical data, explainability becomes an essential and important factor for building trust and supporting the domain experts to validate, critique and refine the model. A novel method for exhibiting explainability have been explored in ProSeNet (Ming et al., 2019). They obtain a prediction by comparing the inputs to few prototype sentences, which are later used to explain a particular decision made by the neural network.

Our main contributions in this project are as follows:

1. We find that averaging word embeddings (Iyyer et al., 2015) is an effective way to obtain the sentence embedding for this task.

2. We develop architectures which involve learnable prototypes and show that the obtained prototype sentences represent typical sentences belonging to each of the classes.

3. We observe that reinforcing the misclassified samples and enriching the prototype layer improves the performance of the model (steerability).

2 Related Work

Historically, traditional systems for sequential sentence classification are based on Naive Bayes (Ruch et al., 2007), support vector machine (SVM)(Liu et al., 2013), Hidden Markov Model (HMM) (Lin et al., 2006), and CRF (Kim et al., 2011; Hassanzadeh et al., 2014). These methods depend on the lexical features such as bag-of-words, semantic features such as the part-of-speech tags, lemmas and statistical features of sentences like sentence positions to perform the classification task. In recent years, they have been replaced with deep-learning approaches involving embedding layers and LSTMs (Hochreiter and Schmidhuber, 1997).

Tagging each sentence in an abstract is significantly different from direct sentence classification such as in sentiment prediction. Context information is critical for sequential sentence classification because the classification of the current sentence could depend on the surrounding sentences or sometimes the entire text. This can be seen as being analogous to POS tagging where words are replaced by sentences and POS tags are replaced by sentence labels. Dernoncourt et al. (Dernoncourt et al., 2016) uses a CRF layer to predict the label of the current sentence depending upon the predicted labels of previous sentences.Jin and Szolovits et al. (Jin and Szolovits, 2018) proposed a hierarchical sequential labeling network to make use of the contextual information within surrounding sentences to help classify the current sentence in medical abstracts. The hierarchical sequential labeling network is built using a token embedding layer , a sentence encoding layer (consisting of CNN or bi-RNN) and a context enriching bi-LSTM layer. This model outperformed the previous state-of-the-art results on two datasets PubMed RCT and NICTA-PIBOSO, for sentence classification in medical abstracts.

Explainabilty of deep learning models has recently gained interest and is crucial for application in fields such as medicine, finance etc where we need to justify the rationale behind the decision. There are three broad approaches to explain deep learning models - treating the deep-learning model as a black-box and training a secondary model to interpret the output for each input such as Lime (Ribeiro et al., 2016), using a gradient based approach such as (Simonyan et al., 2013) or having trainable prototypes as a part of the deep learning model such as (Ming et al., 2019) and (Gee et al., 2019), which makes the model readily interpretable. A prototype here is an embedding vector of a few "typical" examples for each classes. The model first computes the embedding of the input, then the similarities of the input embedding with the prototypes is computed and fed into a feedforward network to classify the embedding. These prototypes are not fixed but are trainable parameters. Post training, a prototype is associated to the input example whose embedding vector is closest to the prototype.

3 Proposed Model

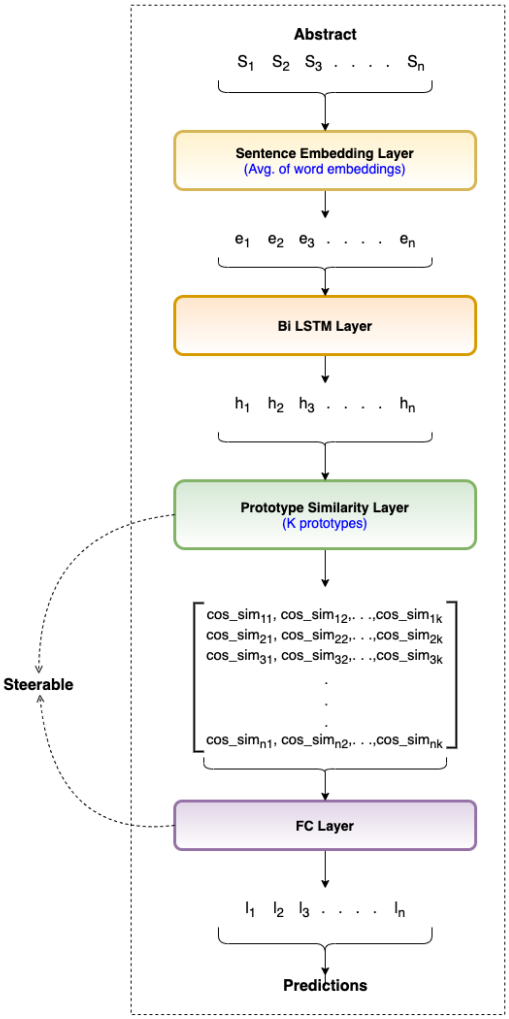

Figure 1: Model architecture

We propose a model architecture which is trained end-to-end, is explainable and steerable. The model consists of the following components:

- **Sentence Embedding Layer**: We used the

average of the word embeddings in a sentence as the sentence embedding (Iyyer et al., 2015). Although this seems to be a very naive approach, we find that it is highly effective. We also tried using a non-linear projection of the sentence embeddings but this did not help the performance. We found that appending the sentence embeddings of the previous and next sentences with the current sentence marginally improves our result.

- **LSTM layer**: We use a bi-directional LSTM to process the sequence of sentence embedding in an abstract. We use the corresponding hidden state of the LSTM as the final representation of a sentence h_i.
- **Prototype Similarity Layer**: Before classifying the above obtained h_i, we pass it through a prototype similarity layer. The prototype similarity layer contains K prototypes each of which are of the same dimension as h_i. Hence, for each sentence we compute a K dimensional vector which is the cosine similarities of the hidden state (h_i) corresponding to the sentence with all the K prototypes. These prototype vectors are randomly initialized and are trained end-to-end.
- **FC Layer**: The final Fully Connected (FC) layer is linear layer which uses the cosine similarities from the prototype layer to predict the class labels.

Once the model is trained, we associate each of the prototypes with a "prototype sentence" which is a sentence in the validation set whose corresponding hidden state is closest to the prototype. Since the final layer is a linear mapping between the cosine similarities and the labels, we can conclude that prototypes which have a high positive weight corresponding to a particular label l_i will positively influence the prediction of that label. Hence, once a label is predicted, we can trace which are the prototypes that positively influenced this prediction and infer that the sentence is similar to the corresponding prototype sentences in the latent space.

Another interesting use of the prototypes is the steerable property. This property described in (Ming et al., 2019) allows experts to add samples which the model is mis-classifying, the corresponding representation of that example will then be added to the list of prototypes and the final linear of the model will be retrained while keeping all the other components frozen. We use this as a form of boosting in our approach, where we randomly select a few examples from the validation set belonging to each of the classes which are misclassified and add their corresponding representations (the LSTM hidden state) to the prototype layer and retrain just the FC layer.

4 Results and Discussion

4.1 Datasets

We use the following datasets on medical scientific abstracts - PubMed RCT 20K and NICTA-PIBOSO. We primarily focus on PubMed RCT 20K dataset as the NICTA dataset contains only 1k abstracts. Table 1 summarizes the statistics of the datasets.

1. **PubMed RCT** This dataset is based on the PubMed database of biomedical literature and is currently the largest dataset for sequential sentence classification. In this dataset, each sentence of each abstract is labeled with their role in the abstract using one of the following classes: background, objective, method, result, or conclusion. Table 2 presents an example abstract comprising structured sentences with their annotated labels.

2. **NICTA-PIBOSO** This dataset was presented as part of ALTA 2012 Shared Task (Amini et al., 2012), to facilitate building automatic sentence classifiers that can classify the sentences from biomedical abstracts into a predefined set of categories for Evidence-Based Medicine (EBM).

4.2 Results

Our primary results are illustrated in Table 3. The metric used here is the weighted F1 score on the test set which is same as the metric used by (Jin and Szolovits, 2018). The state-of-the-art model for this task is (Jin and Szolovits, 2018) which uses a hierarchical-LSTM based approach. Recent approaches using SciBert (Beltagy et al., 2019) do not surpass (Jin and Szolovits, 2018), thus indicating that using BERT may not be useful in this task.

As described in the previous section, we use the average of the word embeddings to obtain the sentence embedding like DAN(Iyyer et al., 2015).

Dataset	Train	Validation	Test	No. of Labels
PubMed	15k(180k)	2.5k(30k)	2.5k(30k)	5
NICTA-PIBOSO	720(7.7k)	80(0.9k)	200(2.2k)	6

Table 1: Total number of abstracts used in training, validation and testing. The value inside the bracket denotes the total number of sentences.

Labels	Sentences
Background	Parathyroidectomy in patients with hyperparathyroidism can produce subsequent increases in bone mineral density (BMD). Ronacaleret, a selective calcium-sensing receptor antagonist that stimulates endogenous parathyroid hormone release, induced mild hyperparathyroidism.
Objectives	The aim of this study is to evaluate whether BMD changes after cessation of ronacaleret treatment.
Methods	Subjects were treated with ronacaleret 100mg (n=16), 200mg (n=38), 300mg (n=35), or 400mg (n=32) once daily, alendronate 70mg (n=17) once weekly, or matching placebo (n=33) for 10-12months; BMD was measured after discontinuation of ronacaleret or alendronate treatment.
Results	At the lumbar spine, all doses of ronacaleret resulted in gains in BMD while on treatment. These increases in BMD were maintained or increased after discontinuation of ronacaleret. All doses of ronacaleret caused bone loss at the total hip while on active treatment. However, there was an attenuation of this loss in the off-treatment extension study.
Conclusions	The gain in BMD at the lumbar spine was maintained post-treatment and the loss of BMD at the total hip was attenuated. We hypothesize that there may have been some bone remineralization after cessation of ronacaleret.

Table 2: Sample classified abstract from the Pubmed 20k dataset.

Before, trying this approach we tried using an LSTM to obtain the sentence embeddings, however we found that the training would then take a very long time, thus limiting the hyperparameter sweeps and experiments, hence we used the DAN approach (denoted as SE_1) in the result tables. We can see that this approach SE_1 + LSTM obtains an F1 score of 89.8 on Pubmed 20k which is better than the score reported using BERT, while having much fewer parameters.

As illustrated in Table 3, we find that using a CRF on top of the LSTM marginally improves the results, we get an F1 score of 90.4 on the Pubmed 20k dataset. However, using a CRF is not directly compatible with a prototype layer, hence we did not go ahead with it. Introducing a prototype layer before the FC layer does not change the F1 score much but makes the model more interpretable as explained earlier. Using a sentence embedding SE_2 which is the concatenation of the average of the word embeddings in the past sentence, current sentence and the next sentence improves the F1 score by 1% which is quite significant. We then exploit the steerable property of the prototypes to boost the model using the validation set. We add the representations of 40 sentences belonging to each of the 5 classes to the prototype layer and retrain the final FC layer. This again gives a significant improvement in the F1 score and we achieve an F1 score of 92.25 which is close to the state-of-the-art (Jin and Szolovits, 2018) which is 92.6. We can surpass the state-of-the-art performance if we ensemble a few of our models trained with different hyper-parameters, however this wont be interpretable.

We also present a study (Table 4), where we compare two strategies to initialize the word embeddings - the first strategy is to use the pretrained embeddings which are trained on the wikepedia-pubmed-pmc texts (we have used this strategy for all the previous results reported), and the second strategy is to randomly initialize the embeddings. We expected the second strategy to do better because using random embeddings meant that we could use embedding of higher dimensions such as $256, 512, 1024$ etc are not limited to using the 200 dimensional pre-trained

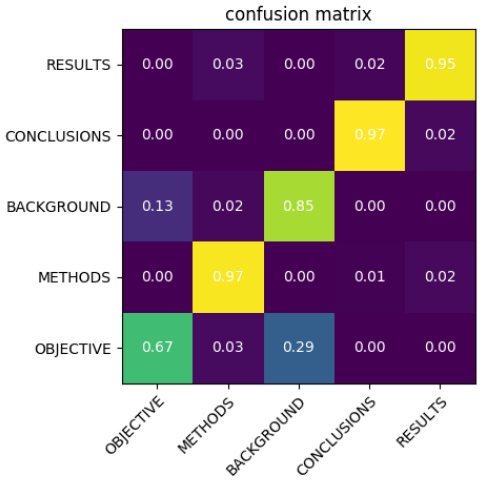

Figure 2: Confusion Matrix obtained by our ensembled model on the Pubmed 20k dataset

embeddings. However, we find that this does not lead to a significant improvement although we are increasing the number of parameters. It is interesting to note that the performance using the random embeddings is less than the pre-trained embeddings by 1% F1 score in the SE_2 case.

We present a few "prototypes sentences" i.e. the sentences associated with the trained prototypes in Table 5. We can see that these are very typical of their respective class. For example, we can see that both objective prototype sentences begin with "To (verb) .. ". We can see that the background prototype sentences have a reference to the past. The result prototype sentences have phrases such as "we compared ..", "demonstrated" etc.

A major problem which we faced was that the models are getting confused between the objective and background classes which is pulling down the F1 score. The confusion matrix in Figure 2 illustrates the confusion between the objective and background classes. When we went through the data, we ourselves found that the distinction between these two classes is not very concrete which is the same reason the models are also facing the trouble with this classification. Some of the misclassified sentences are presented in Table 6. We tried addressing this by two approaches -

- Using a weighted loss function which gives a higher weight to these two classes.
- Sampling more from these two classes while steering.

We could not get any significant improvement with both these techniques but the second technique marginally improved the result (by around 0.02).

Model	PubMed	NICTA
Best Published Models		
SciBert (Beltagy et al., 2019)	86.81	-
State of the art(Jin and Szolovits, 2018)	92.60	84.30
Our Models		
SE_1 + LSTM + FC	89.85	78.18
SE_1 + LSTM + CRF + FC	90.40	79.90
SE_1 + LSTM + ρ + FC	89.92	78.10
SE_2 + LSTM + ρ + FC	90.90	80.45
SE_2 + LSTM + ρ + FC + Steering	92.25	82.80
Ensembled model	92.72	83.35

Table 3: Comparison of F1 scores of our model with existing state of the art models. SE_1 is the sentence embedding layer where we use just that particular sentence to compute the embedding. SE_2 is the sentence embedding layer where in the embeddings of one previous & one next sentence are concatenated. ρ is the prototype layer. FC denotes the fully-connected layer. The token embeddings obtained from the pretrained wikepedia-pubmed-pmc texts were used to get the sentence embeddings.

Model	Pre-trained	Random
SE_1 + LSTM + FC	89.85	88.28
SE_1 + LSTM + ρ + FC	89.92	89.96
SE_2 + LSTM + ρ +FC	90.90	89.94

Table 4: Comparison of F1 scores of our model for the PubMed dataset using pretrained embeddings from wikepedia-pubmed-pmc texts vs random embeddings of dimension 512. Definition for SE_1, SE_2, FC and ρ are provided in the caption of Table 1.

4.3 Experimental Details and Hyper parameters

Here, we list some of the finer details which will be helpful for the reproduction of this work:

- We performed early stopping - used the state of the model when it achieved the best validation score to obtain the test metrics.
- We found that the token embeddings which are pre-trained on a large corpus of Wikipedia, PubMed and PMC texts are better than using random embedding.

Labels	Prototype Examplar Sentences
Background	Despite current screening and therapeutic options (pre-biologics) @ % to @ % of children with JIA-associated uveitis may develop bilateral visual impairment and certified legally blind . Previous studies have shown that a calcium (Ca) pre-rinse given before a @ ppm fluoride (F) rinse greatly increased salivary fluoride .
Objectives	To compare efficacy and safety of thromboprophylaxis with semuloparin started post-operatively versus enoxaparin started preoperatively in major abdominal surgery. To evaluate effects of preoperative high-dose glucocorticoid on the inflammatory response and recovery after endovascular aortic aneurysm repair (EVAR).
Methods	This study provides Class I evidence that for patients with inclusion body myositis , bimagrumab increases thigh muscle volume at @ weeks. Patients who have a malignant condition may have a significantly higher risk of developing NLDVT , and patients with NLDVT , compared with those without , appeared to be at higher risk of PE but not higher risk of death .
Results	We compared PKEP with OP for large prostates and found that PKEP is less invasive, with short - to long-term micturition improvement equivalent to OP. Both groups demonstrated significant end-of-treatment cessation rates .
Conclusions	This feasibility trial identified adherence and follow-up rates and sample-size estimates important to the conduct of a fully powered efficacy trial. In situ uterine closure is more advantageous than extra-abdominal repair in terms of surgery duration, postoperative pain and need for additional analgesia, and return of bowel movement.

Table 5: Examples of prototype sentences for each of the five classes of the PubMed dataset

Predicted	Truth	Sentence
Background	Objective	The aim of this study is to evaluate the effect of GLP-@ analogue liraglutide on glycaemic control in patients with type @ diabetes treated with MDI with inadequate glycaemic control. This study assessed whether diets with different fat quality and supplementation with coenzyme Q@ (CoQ) affect the metabolomic profile in urine analyzed by proton nuclear magnetic resonance spectroscopy from elderly people.
Objective	Background	The aim of this study is to evaluate the effect of GLP-@ analogue liraglutide on glycaemic control in patients with type @ diabetes treated with MDI with inadequate glycaemic control. In the present study, we investigated the effects of HQH on patients with mild immunoglobulin A nephropathy (IgAN) through a prospective randomized controlled study.

Table 6: Examples of incorrect predictions performed by our model.

- Freezing the embedding layer after one epoch is an effective form of regularization.
- We used the Adam optimizer with an initial learning rate of 0.003 and decay it by a factor

of 0.9 after every epoch.

- We used a batch size of 16 for the Pubmed dataset and a batch size of 1 for the NICTA dataset and we shuffle the training samples for each epoch to avoid overfitting.

5 Conclusion

In this work, we presented an interpretable deep-learning model for sentence classification of medical abstracts which obtains a near state-of-the-art performance. We show that using prototypes for this task can make the models readily interpretable without any loss in performance. We also observe that DAN is surprisingly powerful as reported by many other works. We observe that the prototype sentences indeed represent the typical way a sentence belonging to a particular class would look like.

6 Future Work

We are currently working on refining the prototype sentences to prototype phrases which would further aid the interpretability of the model.

Acknowledgement

We thank Prof. Greg Durrett for the helpful discussions and constructive feedback.

References

Iman Amini, David Martinez, Diego Molla, et al. 2012. Overview of the alta 2012 shared task.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: Pretrained language model for scientific text. In *EMNLP*.

Franck Dernoncourt and Ji Young Lee. 2017. Pubmed 200k rct: a dataset for sequential sentence classification in medical abstracts. *arXiv preprint arXiv:1710.06071*.

Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. 2016. Neural networks for joint sentence classification in medical paper abstracts. *arXiv preprint arXiv:1612.05251*.

Alan H Gee, Diego Garcia-Olano, Joydeep Ghosh, and David Paydarfar. 2019. Explaining deep classification of time-series data with learned prototypes. *arXiv preprint arXiv:1904.08935*.

Hamed Hassanzadeh, Tudor Groza, and Jane Hunter. 2014. Identifying scientific artefacts in biomedical literature: The evidence based medicine use case. *Journal of biomedical informatics*, 49:159–170.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1681–1691.

Di Jin and Peter Szolovits. 2018. Hierarchical neural networks for sequential sentence classification in medical scientific abstracts. *arXiv preprint arXiv:1808.06161*.

Arif E Jinha. 2010. Article 50 million: an estimate of the number of scholarly articles in existence. *Learned Publishing*, 23(3):258–263.

Su Nam Kim, David Martinez, Lawrence Cavedon, and Lars Yencken. 2011. Automatic classification of sentences to support evidence based medicine. In *BMC bioinformatics*, volume 12, page S5. BioMed Central.

Jimmy Lin, Damianos Karakos, Dina Demner-Fushman, and Sanjeev Khudanpur. 2006. Generative content models for structural analysis of medical abstracts. In *Proceedings of the hlt-naacl bionlp workshop on linking natural language and biology*, pages 65–72. Association for Computational Linguistics.

Yuanchao Liu, Feng Wu, Ming Liu, and Bingquan Liu. 2013. Abstract sentence classification for scientific papers based on transductive svm. *Computer and Information Science*, 6(4):125.

Yao Ming, Panpan Xu, Huamin Qu, and Liu Ren. 2019. Interpretable and steerable sequence learning via prototypes.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.

Patrick Ruch, Celia Boyer, Christine Chichester, Imad Tbahriti, Antoine Geissbühler, Paul Fabry, Julien Gobeill, Violaine Pillet, Dietrich Rebholz-Schuhmann, Christian Lovis, et al. 2007. Using argumentation to extract key sentences from biomedical abstracts. *International journal of medical informatics*, 76(2-3):195–200.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Mark Ware and Michael Mabe. 2015. The stm report: An overview of scientific and scholarly journal publishing.