Neural language models for text classification in evidence-based medicine

Andrés Carvallo (afcarvallo@uc.cl)[1], Denis Parra [1],Gabriel Rada[*] , Daniel Pérez[2], Juan Vásquez[2], Camilo Vergara[2]

1. Pontificia Universidad Católica de Chile.
2. Epistemonikos Foundation.

Millennium Institute Foundational Research on Data

Epistemonikos

Motivation

- Revision of medical evidence is increasingly difficult due to the exponential increase of available evidence since the COVID-19 pandemic started—more than 200,000 new articles with 1,000 articles growth everyday.
- Medical evidence databases on the web are PubMed, Medline, among others.
- Epistemonikos, our main collaborator, is a foundation in charge of curating evidence with physician's help.

Introduction

- Artificial Intelligence has many medical applications such as image classification, automatic analysis of medical records, and evidence-based medicine.
- Evidence-based medicine seeks to support medical decisions with scientific evidence. (Berwick et al., 2008)
- Search engines such as PubMed or Medline are an alternative but have many data articles (graph 1) that are not validated by health experts.
- Epistemonikos Foundation has more than 500 experts who validate that each document is relevant to medical issues, but it can take a lot of effort and time (Bekhuis et al., 2014).
- In this proposal, the use of artificial intelligence is proposed to facilitate experts in this process of review and classification of relevant evidence depending on the article's type of study methodologies and medical relevance.

Objectives

- Improve the efficiency of document search in the practice of evidence-based medicine.
- Facilitate the classification of study methodologies and types of relevant documents for evidence-based medicine.

Dataset

Systematic Reviews: 286,050

Epistemonikos

Primary Studies Non-RCT: 35,644

Primary Studies RCT: 56,623

Broad Synthesis of a Systematic review: 17,320

Excluded 6,096

Methodology

To identify if the article is relevant evidence for COVID-19 or other disease physicians evaluate:

1. Study Methodology
   a. Randomized Controlled Trial

2. Type of article:
   a. Systematic Review
   b. Primary study using randomized controlled trial.

We propose an "XLNET transformer-based language model (Yang et al, 2019)" that according to the content of the title and abstract of the article:

Predicts type and study methodology.

Decide to include it or not in the Epistemonikos Database.

Then compare this model with other baselines:

1. Random Forest (current solution by Epistemonikos): which uses a tokenizer based on the appearance of keywords to classify documents.

2. BioBERT (Lee et al, 2020): BERT (Devlin et al, 2018) language model pre-trained on PubMed full text articles.

Results and Discussion

		Random Forest			XLNet			BioBERT		
	# docs.	Prec.	Rec.	F-1	Prec.	Rec.	F-1	Prec.	Rec.	F-1
Broad synthesis	17,324	.75	.15	.26	.67	.56	**.61**	0	0	0
Systematic review	286,050	.93	.99	.96	.96	.98	**.97**	.85	1.0	.92
Primary rct	56,623	.25	.79	.38	.94	.85	**.89**	.71	.71	.71
Primary non-rct	35,644	.63	.40	.49	.64	.91	**.75**	.61	.90	.72
Excluded	6,096	.70	.21	.32	.82	.74	**.78**	0	0	0

Table 1. Results obtained from XLNET for document classification compared to other baselines in terms of precision, recall and f1-score.

We can observe that in terms of F1 score which is the harmonic mean of precision and recall XLNET makes an improvement over BioBERT and Random Forest.

Figure 1. . U-MAP dimensionality reduction of XLNET representation fo medical articles for each class. .

We made a U-MAP dimensionality reduction of documents represented with XLNET and it can be seen that types of documents are well separated.

Conclusions

- XLNET showed a considerable improvement over baselines. They are even identifying minority classes, that although they are not relevant evidence, an expert needs to take into consideration.
- With this proposed solution, we are, facilitating the identification of relevant evidence for physicians, especially in COVID-19 pandemic times, were articles on the web have been exponential.
- There is still space for improvement, for example, considering other features from articles.

References

BEKHUIS,T, TSEYTLIN, E, MITCHELL, K. J., & DEMNER-FUSHMAN D. 2014. Feature engineering and a proposed decision-support system
BERWICK, D. M., NOLAN, T. W., & WHITTINGTON, J. (2008). The triple aim: care, health, and cost. Health affairs.
YANG, Z.,DAI, Z., YANG, Y., CARBONELL, J., SALAHUDINOV, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pre-training for language understanding. In Advances in neural information processing systems Lee, J.,YOON, W., KIM, S., KIM, D., KIM, S., SO, C. H., & KANG, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics,

Discussion and future work

One essential point of discussion is why XLNET outperforms the other baselines.

- We hypothesize that one of the possible reasons is that XLNET allows you to use all the content of the document without having to cut it into 512 characters like BERT.
- Another reason is that XLNET uses improved training methodology (permutation language modeling, where all tokens are predicted but in random order).

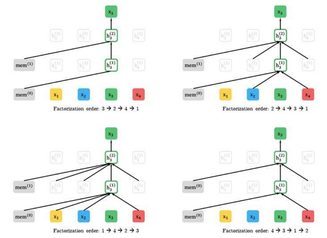

Figure 1: Illustration of the permutation language modeling objective for predicting x_3 given the same input sequence x but with different factorization orders.

Figure 2. XLNET permutation language modeling process.

For future work we will implement a user study to:

- Evaluate if explanations are relevant for physicians when using automated models.

- Compare different interfaces and visual encodings.

Figure 3. User study prototype to evaluate the effect of explanations in physicians for evidence based medicine.

Acknowledgments

- Pontificia Universidad Catolica of Chile
- Millenium Institute Foundational Research on Data.
- Epistemonikos Foundation.