Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230252

Health-Related Content in Transformer-Based Language Models: Exploring Bias in Domain General vs. Domain Specific Training Sets

Giuseppe SAMO[a,b,1] and Caterina BONAN[c]
[a] *University of Geneva, Switzerland*
[b] *Beijing Language and Culture University, China*
[c] *University of Cambridge, UK*
ORCiD ID: Giuseppe Samo https://orcid.org/0000-0003-3449-8006,
Caterina Bonan https://orcid.org/0000-0002-4808-6865

Abstract. In this communication, we demonstrate that the bias observed in domain general training sets with health-related content is not improved in domain specific health-communication corpora, contra.

Keywords. Natural Language Processing, Health-content, Language Models, Knowledge Reproduction, Corpora, COVID-19

1. Introduction

Artificial neural language models (LMs) parse and generate complex linguistic structures across languages [1,2], and can be used for fact checking [3]. In [1], we detected syntactic bias in Transformer-based LMs using a list of myth busters on COVID-19 from parallel World Health Organization corpora. We demonstrated that when LMs are queried with sentences compared with ad hoc examples with opposite polarity, asymmetries are easily detected and quantified. Here, we explore six training sets for English and Chinese LMs to detect bias with respect to domain-specific (i.e., medical) semantico-encyclopedic knowledge, adopting and improving the dataset previously discussed in [1].[2]

2. The study

Hypothesis: In [1]'s conclusions, we tentatively attributed the observed bias to the type of training data under investigation, and hypothesised that LMs trained on domain specific datasets might perform better than those trained with domain general data.

[1] Corresponding Author: Giuseppe Samo, Giuseppe Samo, Department of Linguistics, University of Geneva, mail: Rue de Candolle 2, 1205 Geneva, Switzerland. E-mail: giuseppe.samo@unige.ch.

[2] Information on the data is available at the following link:https://github.com/samo-g/health-transformer.

Materials: The queried parallel datasets for English and Chinese are those presented in The training sets for this contribution are presented in table 1.
Methods: As in [1], the machine is presented with two sentences: (i) a TRUE statement from the corpus and (ii) a FALSE counterpart created using logical operators. Our measure is represented by the difference between the surprisal of the TRUE and the FALSE statement. The surprisal is the logarithm of the reciprocal of the output probability in a fill-mask task. In a nutshell, lower surprisal is to be expected for TRUE sentences.

Table 1. Language Models used in this paper. All references are available as hyperlinks.

Languages	Training Set Type
English	Domain general (web, wiki): BERT, BigBird
	Domain specific: HealthF (fact checking), PHS (health surveillance on social media)
Chinese	Domain general (web): Bert-base-Chinese
	Domain specific: HealthZH (medical dialogues, records, textbooks)

3. Results and Conclusions

Figure 1 displays an asymmetry for English ($F(3, 64) = 177.30362$; $p < .00001$): while the BERT general domain LM performs worse (M = 18.3, SD = 5.82) than domain specific ones, BigBird (M = -24.5, SD = 7.08) shows the least surprisal on TRUE statements, plausibly due to its architecture combining sparse and global attention.

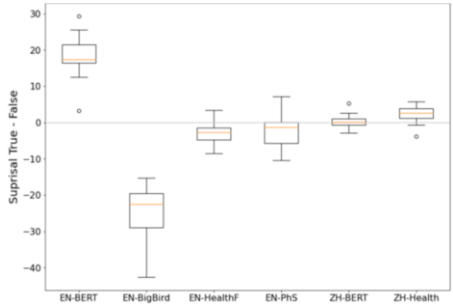

Figure 1. True-False across language models.

In Chinese, the general domain performs better ($M = 0.51$, $SD = 2.42$) than the domain-specific one ($M = 1.81$, $SD = 2.42$) ($t(34) = 2.6837$, $p < .05$). [1] provided a methodology for the detection of bias in LMs that can be adopted to fact checking and the general medical domain, and predicted lower bias with health-specific training sets. Here, we showed that domain specific training sets do not necessarily perform better than domain general ones. Future studies should explore additional datasets (e.g., online medical training, [4]) and languages.

References

[1] Samo G, Bonan C, Si F. Health-Related Content in Transformer-Based Deep Neural Network Language Models: Exploring Cross-Linguistic Syntactic Bias. Stud Health Technol Inform. 2022;295:221-225.
[2] Linzen T, Baroni M. Syntactic Structure from Deep Learning. Annual Review of Linguistics 7:1, Jan.2021;195-212.
[3] Ahmed S, Hinkelmann K, Corradini F, Development of fake news model using machine learning through natural language processing; arXiv:2201.07489, Jan. 2022.
[4] Utunen H, et al, Global Reach of an Online COVID- 19 Course in Multiple Languages on OpenWHO in the First Quarter of 2020: Analysis of Platform Use Data, J Med Internet Res. 22 (2020) e19076.