Advances in Informatics, Management and Technology in Healthcare
J. Mantas et al. (Eds.)

doi:10.3233/SHTI220702

Health-Related Content in Transformer-Based Deep Neural Network Language Models: Exploring Cross-Linguistic Syntactic Bias

Giuseppe SAMO[a,1], Caterina BONAN[b] and Fuzhen SI[a]
[a] *Beijing Language and Culture University*
[b] *University of Cambridge*

Abstract. This paper explores a methodology for bias quantification in transformer-based deep neural network language models for Chinese, English, and French. When queried with health-related mythbusters on COVID-19, we observe a bias that is not of a semantic/encyclopaedical knowledge nature, but rather a syntactic one, as predicted by theoretical insights of structural complexity. Our results highlight the need for the creation of health-communication corpora as training sets for deep learning.

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

1. Introduction

The outbreak of COVID-19 has highlighted the need of local and global digital solutions to provide fast and effective multilingual knowledge-transfer, such as health responses in emergency contexts [1]. Recent developments in computational linguistics have demonstrated the ability of artificial neural network architectures to parse complex linguistic structures cross-linguistically [2], and the interest of implementing BERT transformer architectures [3] in Neural Machine Translation (see [4] for a detailed discussion).

From the point of view of formal linguistics, one of the advantages of using BERT as a language model [5] is that these architectures can be investigated to detect forms of lexical, morphological, syntactic, or semantic (at clausal level) bias. These models, when queried with naturally occurring sentences compared with devised *ex-novo* examples differing in at least one lexical/syntactic variable, allow us to elegantly quantify asymmetries between pairs of isolated conditions.

[1] Corresponding Author, Giuseppe Samo, Department of Linguistics, Beijing Language and Culture University, Mailbox 82, Xue Yuan Road 15, 100083 Beijing, People's Republic of China; E-mail: samo@blcu.edu.cn.

In this paper, we perform a study on syntactic bias quantification utilising a set of health-related mythbusters published on the World Health Organization's website.[2] Our goal is to observe whether these language models are able to mark naturally occurring examples (i.e., the sentences extracted from the mythbuster repository) from devised *ex-novo* sentences that differ from their original counterparts merely by the addition/substraction of a syntactic negation. Conveniently, we label the original sentences as *true*, and the latter as *false*.

The novelty of this work, beyond the methodology, is its crosslinguistic nature. We will indeed work on three languages belonging to three different language families, namely Chinese, English, and French. The model for syntactic bias quantification proposed in this contribution, with its current limits and dimensions of improvement, can contribute to the success of existing approaches to fact-checking (see [6], and reference therein), especially in health-related content.

In the following, Section 2 briefly introduces language models. Section 3 presents the materials, methods, and the results of our study. In section 4, we discuss some desirable further improvements to the model presented and conclude.

2. Exploring language models to quantify bias

When provided with a linguistic context, large contextual word embedding, deep multi-layer models such as BERT [3] are able to predict the conditional probability of a next word in input. BERT is bidirectional and thus endowed with better predictive power than unidirectional architectures. In this paper, we explore a measure of surprisal, the logarithm of the reciprocal of this probability ([5] for a detailed discussion, applications, and references). BERT models are trained on general domain large-scale corpora such as news or encyclopedic entries. Domain specific BERT models do exist, such as FinBERT [7], trained with financial communication corpora, or ClinicalBERT, [8], with clinical notes.

Our primary interest is the parsing ability of language models in health-related knowledge-production and communication content [9], which show similarities with encyclopedic entries (see the metrics in [10]). Specifically, two comparable syntactic structures are given as input to the transformer: (i) one naturally occurring example extracted from our reference corpus and evaluated as *true* health-related content (mythbuster); (ii) a devised *ex-novo* sentence built with the logical opposite created by adding/removing the syntactic negation (see [11] for theoretical and applied dimensions), which represents a *false* statement.

We define as 'bias' the preference for one or the other syntactic configuration, in different syntactic regions (see section 3), in terms of a lower surprisal between the *true* and the *false* statements. We also establish a coefficient (labelled here as *T-F*), given by the difference between the surprisal outputted for a *true* configuration minus the surprisal for the *false* counterpart. In line with the model developed and discussed in [10], we explore a restricted corpus of sentences in the three languages of choice. If adequately demonstrated and improved, this model based on formal approaches of syntactic tree-

[2] Chinese: https://www.who.int/zh/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters (03/2022); English: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters (03/2022); French: https://www.who.int/fr/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters (03/2022).

structures can be developed for automatic text/sentence classification in small-size data and for the exploration of fine-grained research questions.

3. The study

*Materials:*Our input sentences were extracted from the mythbusters webpage of the World Health Organization (footnote 2, the content partiallty varies among languages). A total of 51 *facts* (18 Chinese, 17 English, and 16 French) for a total of 102 constructions were manually coded and divided in syntactic regions, as in Table 1. We created a second 'dependency region' for 8 Chinese sentences to host complements preceding the verb/predicative structure (marked as Dep' in the results).[3]

Table 1. Examples of the experimental items (Lang = Languages, Val = Value, Pol = Syntactic Polarity, Subj = Subject, Pred = Predicative structure, Dep = other dependents (such as objects/adjuncts).

Lang	Val	Pol	Subj (Region 1)	Pred (Region 2)	Dep (Region 3)
En	True	Pos	*Hand sanitisers*	*can be used*	*often*
En	False	Neg	*Hand sanitisers*	*cannot be used*	*often*
Fr	True	Neg	*Manger de l'ail*	*n'aide pas*	*à prévenir la COVID-19*
Fr	False	Pos	*Manger de l'ail*	*aide*	*à prévenir la COVID-19*
Zh	True	Neg	低温和冰雪	不能杀死	COVID-19 病毒。
Zh	False	Pos	低温和冰雪	能杀死	COVID-19 病毒。

Methods: We adopt the pre-trained model of type transformer BERT for English (bert-large-cased, [3]), Flau-BERT for French (flaubert_base_cased, [11]) and Chinese BERT for Chinese (bert-base-chinese).[4] The models output a surprisal measure in the context of a *fill-mask* task that hides target words from the region of a structure.[5]

4. Results

Results are given in Table 2 and Figure 1.

Table 2. Preferences (lower surprisal) in *true* conditions. *Regions in both conditions had the same surprisal.

	Chinese (Zh)				English (En)			French (Fr)		
	Subj	**Dep'**	**Pred**	**Dep**	**Subj**	**Pred**	**Dep**	**Subj**	**Pred**	**Dep**
Preferred *true*	6/18	6/8	1/18*	5/18	1/17	0/17	0/17	9/16	6/16	10/16
If *true* **positive**	2/3	1/1	1/3*	3/3	1/8	0/8	0/8	3/4	4/4	3/4
If *true* **negative**	4/15	5/7	0/15*	2/15	0/9	0/9	0/9	6/12	2/12	7/12

[3] We only tested monoclausal structures and non-quantified predicates. We also removed the cases of post-verbal subjects in French. We also carefully uncased capital letters in English and French.

[4] https://huggingface.co/bert-base-chinese (03/2022).

[5] All the experimental stimuli and relevant materials can be found at https://github.com/samo-g/health-transformer (04/2022).

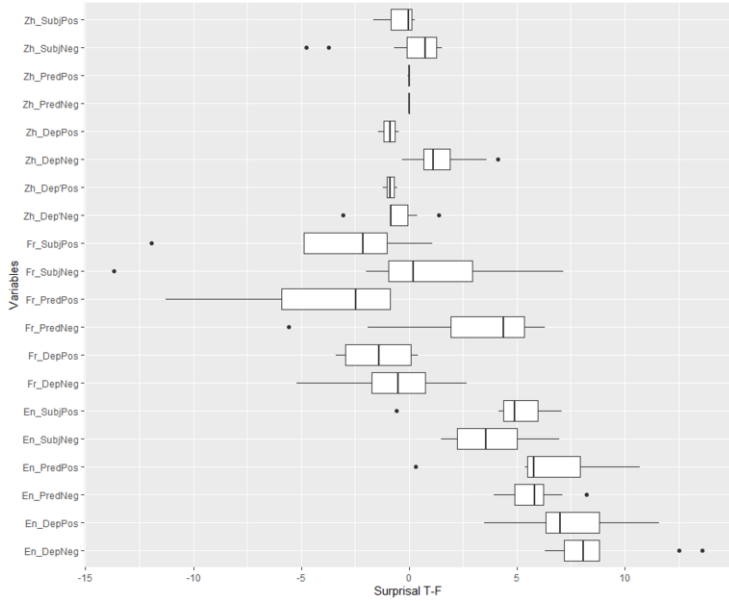

Figure 1. Languages, regions, and coefficient *T-F.* Negative *T-F* indicates preferences for the *true* conditions. 0 is the threshold of preference.

The results in Table 2 demonstrate that, in the three languages under investigations, lower surprisal does not correlate with a *true* in our set of constructions, implying that semantic and encyclopedic knowledge is plausibly not encoded. We observe, in this respect, a slight preference for positive (i.e., non negated) constructions in French (*binomial p = .016*), and marginally also in Chinese (*binomial p = 0.193*), but clearly not in English (*binomial p = <0.01, z = -5,5, p < 0.01*). On the other hand, the manipulation of the negation on Chinese predication does not show clear asymmetries. Figure 1 presents the details of coefficient $T - F$, calculated as the difference between the surprisal for *true* and *false* constructions. Figure 1 highlights that a syntactic bias emerges only in French positive sentences ($t(46)= 2.90522$, $p < .01$). This result is expected in light of recent understandings of syntactic complexity in knowledge-transfer ([13]), as the French negation involves two morphosyntactic elements. Finally, our results show that there is no clear mapping between semantic/encyclopaedical knowledge and surprisal, at least with respect to our dataset of mythbusters.

5. Conclusions

We presented and tested a methodology for quantifying bias of a training set, in which the only bias found is of syntactic nature, whereas semantic/encyclopaedical knowledge was not detected. An additional result is related to the syntactic complexity of split negations, as observed for French, whose theoretical intricacies exceed the scope of this contribution. Finally, not only does our study highlight the need to create health-communication corpora as training sets for deep learning, but will also have the potential to help measure and improve the quality of such corpora, for example in addressing miscommunication across the globe in critical health interventions (e.g., vaccination).

References

[1] Utunen P, Ndiaye N, Attias M, Mattar L. Multilingual Approach to COVID-19 Online Learning Response on OpenWHO.org. In: Mantas J, Hasman A, Househ MS, Gallos P, Zoulias E, Liaskos J, editors. Informatics and Technology in Clinical Care and Public Health; IOS Press, 2022;Jan:192-5.

[2] Linzen T, Baroni M. Syntactic Structure from Deep Learning. Annual Review of Linguistics 7:1, Jan.2021;195-212.

[3] Devlin J, Ming WC, Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Burstein J, Doran C, Solorio T, editors. Proceedings of the 2019 Conference of the North American Chapter of the ACL: Human Language Technologies; 2002 Jun 2-7; Minneapolis, Minnesota. Association for Computational Linguistics; p. 4171–86.

[4] Shavarani HS, Sarkar A. Better Neural Machine Translation by Extracting Linguistic Information from BERT. In: Merlo P, Tiedemann J, Tsarfaty R, editors. Proceedings of the 16th Conference of the European Chapter of the ACL, 2021 Apr 19 – 23; Online, Association for Computational Linguistics; p. 2772–83.

[5] Hale J. Information-theoretical complexity metrics. Language and Linguistics Compass (10), 2016, Aug. 397– 412.

[6] Ahmed S, Hinkelmann K, Corradini F, Combining Machine Learning with Knowledge Engineering to detect Fake News in Social Networks-a survey. In: Martin A, Hinkelmman K, Gerber A, Lenat D, van Harmelen F, Clark P, editors. Proceedings of the AAAI 2019 Spring Symposium on Combining Machine Learning with Knowledge Engineering; Mar 25-27; Palo Alto, CA; University of Stanford; p. 1 – 12.

[7] Yang Y, Uy MCS, Huang A. Finbert: A pretrained language model for financial communications. arXiv preprint, arXiv:2006.08097, 2020.

[8] Huang K, Altosaar J, Ranganath R. Clinicalbert: Modeling clinical notes and predicting hospital readmission. arXiv preprint, arXiv:1904.05342, 2019.

[9] Zhao Y, Samo G, Utunen H, Stucke O, Gamhewage G. Evaluating Complexity of Digital Learning in a Multilingual Context: A Cross-Linguistic Study on WHO's Emergency Learning Platform. Stud Health Technol Inform 2021 May 27;281:516-517. doi: 10.3233/SHTI210222

[10] Samo G, Zhao Y, Guasti MT, Utunen H, Stucke O, Gamhewage G. Could linguistic complexity be automatically evaluated? A Multilingual Study on WHO's Emergency Learning Platform. In: Mantas J, Hasman A, Househ MS, Gallos P, Zoulias E, Liaskos J, editors. Informatics and Technology in Clinical Care and Public Health; IOS Press, Jan 2022;196-9.

[11] Tettamanti M, Manenti R, Della Rosa PA, Falini A, Perani D, Cappa SF, Moro A. Negation in the brain: modulating action representations. Neuroimage. 2008 Nov 1;43(2):358-67.

[12] Le H, Vial L, Frej J, Segonne V, Coavoux M, Lecouteux B, Allauzen A, Crabbe B, Besacier L, Schwab D. Flaubert: Unsupervised language model pre-training for French. In: Calzolari N, Béchet F, Blache P, Choukri K, Cieri C, Declerck T, Goggi S, Isahara H, Maegaard B, Mariani M, Mazo H, Moreno A, Odijk J, Piperidis S, editors. Proceedings of the 12th Language Resources and Evaluation Conference, 2020 May 11-16, Marseille, France; p. 2479–90.

[13] Samo G, Zhao Y, Gamhewage G. Syntactic Complexity of Learning Content in Italian for COVID-19 Frontline Responders: A Study on WHO's Emergency Learning Platform. Verbum. Dec 2020;11:1-4.