A Multilevel Analysis of PubMed-only BERT-based Biomedical Models

Vicente Ivan Sanchez Carmona and **Shanshan Jiang** and **Bin Dong**
Ricoh Software Research Center (Beijing) Co., Ltd
{Vicente.Carmona, Shanshan.Jiang, Bin.Dong}@cn.ricoh.com

Abstract

Biomedical NLP models play a big role in the automatic extraction of information from biomedical documents, such as COVID research papers. Three landmark models have led the way in this area: BioBERT, MSR BiomedBERT, and BioLinkBERT. However, their shallow evaluation –a single mean score– forbid us to better understand how the contributions proposed in each model advance the Biomedical NLP field. We show through a Multilevel Analysis how we can assess these contributions. Our analyses across 5000 fine-tuned models show that, actually, BiomedBERT's true effect is bigger than BioLinkBERT's effect, and the success of BioLinkBERT does not seem to be due to its contribution –the Link function– but due to an unknown factor.

1 Introduction

Machine reading of biomedical texts has greatly advanced due to pretrained NLP models such as BERT. Biomedical NLP applications are of great value due to their utility in real-world scenarios such as answering questions which require background knowledge or the extraction of complex biomedical entities from astonishing volumes of academic papers related to COVID, for example.

Of special acknowledgement, three BERT-based biomedical models, trained on PubMed abstracts (their only source of biomedical knowledge), led the way to concise research contributions on biomedical NLP, namely, BioBERT (Lee et al., 2019) proposing Domain Adaptive Pretraining (DAPT), MSR BiomedBERT[1] (Gu et al., 2021, which we refer to as BiomedBERT) which challenged DAPT by pretraining BERT from scratch with PubMed abstracts, and BioLinkBERT (Yasunaga et al., 2022) which implemented a way to link hyperlinked documents at pretraining time – the Link function. These 3 contributions resulted in significant improvements on downstream scores on the BLURB benchmark (Gu et al., 2021).

However, we claim, current evaluation methods are oversimplistic. They reduce to a simple mean score across datasets in the BLURB suite –a single estimate. This forbid us to better understand the contributions proposed by each work such as their effect on scores and interaction with downstream datasets. Moreover, from this single estimate, how can we disentangle the contributions' effects from the effects of other variables such as random seeds, learning rates, or number of epochs? We cannot. And while some works show ablation studies to see the particular effects of the proposed contribution, doing so to isolate it from all possible variables (including those mentioned above) leads to an exponential number of ablations which results in a non-environmentally friendly, unfeasible approach if pretraining is necessary for each ablation.

In this paper, we propose a regression analysis widely used in the fields of Psychology and the Social Sciences –Multilevel Analysis– to account for the effects of all measurable variables, without the need for ablations or further pretraining experiments, in order to disentangle their effects from the true effect of the proposed contributions from BioBERT, BiomedBERT, and BioLinkBERT. Our analyses show that, actually, random seeds have a big effect on downstream scores. Also, while BioBERT's and BiomedBERT's contributions have a big and significant effect by improving on vanilla BERT's score by 2.25 and 4.36 points, respectively, on average across BLURB datasets, BioLinkBERT's Link function shows only a big effect for QA datasets but not for any other dataset.

2 Background and Related Work

2.1 Multilevel Regression Analysis

Multilevel models (MLMs) are a type of regression analysis where the outcome to be modeled

[1] Previously known as PubMedBERT.

Proceedings of the 6th Clinical Natural Language Processing Workshop, pages 105–110

(downstream scores in our case) is dependent on a set of independent variables that can pertain to different levels in a hierarchy. In our case, we define our problem as a 2-level hierarchy where the lowest level –Level 1– contains fine-tuned models, which is nested inside the upper level –Level 2– which corresponds to groups of fine-tuned models grouped according to the choice of pretrained model and downstream dataset; for example, BioBERT-BIOSSES is a group of BioBERT models fine-tuned on the BIOSSES dataset.[2]

Thus, variables at level 1 describe fine-tuning attributes such as learning rates, batch size, and number of epochs. On the other hand, level-2 variables describe attributes of the pretrained models, such as the contribution proposed by a work (for example, the Link function proposed by BioLinkBERT), and the choice of downstream dataset. In this way, a 2-level MLM (de Leeuw and Meijer, 2008) can be expressed as:

$$y = \beta_0 + \sum_{fixed} \beta_i x_i + \sum_{random} \gamma_{ij} x_{ij} + u_{0j} + e \quad (1)$$

where β_0 is the grand-mean intercept; the first summation corresponds to level-1 and level-2 *fixed-effects* coefficients (β_i) which represent the average individual effect of each variable (x_i) on downstream scores (y); the second summation is a key term that distinguishes MLMs from other regression models: level-1 *random-effects*, i.e. an *adjusted* effect (γ_{ij}) on the level-1 fixed-effects coefficients according to each group (indexed by j);[3] and similarly for the random intercepts u_{0j} which are adjusted effects for each group on the grand-mean intercept; finally, e is the residual. This model can be fitted using Maximum Likelihood Estimation or variants.

2.2 MLMs for Experimental Analyses

MLMs[4] have been widely used for analysing experimental and observational data by fields such as Psychology (Muradoglu et al., 2023; Judd et al., 2017), Linguistics (Baayen et al., 2008), and the Social Sciences (Rasbash et al., 2010; de Leeuw and Meijer, 2008). For example, in the field of Education, MLMs analyze the impact of both student (level-1) variables (age, socioeconomic status, gender) and school (level-2) variables (mean socioeconomic status, ethnicity proportions) on students' academic performance (Goldstein et al., 2007).

Works in Psychology have used MLMs to disentangle the effects of different variables at different levels while measuring their impact on participants' reaction time on cognitive tasks (Kliegl et al., 2011, 2010). Moreover, work in Linguistics has leveraged MLMs to model the effect of between-speaker features (age, country, etc.) and within-speaker features (length of sentence, sequential position of phrase, etc.) on articulation rate of spoken sentences (Quené, 2008).

To our knowledge, our work is the first approach towards leveraging MLMs for analysis of biomedical NLP models.

3 Dataset and Multilevel Model

3.1 Dataset for Multilevel Analysis

To generate a dataset to fit an MLM that explains the impact of variables on downstream scores, we fine-tune[5] BioBERT, BiomedBERT, BioLinkBERT and vanilla BERT (which we use as baseline) on all datasets in the BLURB suite. We use test set scores as the dependent variable. And we use fine-tuning and pretraining features as level-1 and level-2 variables, respectively.

To obtain robust estimates of effects (regression coefficients) we not only include test scores from the best-validation-score models,[6] we also include the scores from a vicinity around the best-validation-score models. This vicinity is defined around the values of variables that lead to the best validation score, (namely learning rate, batch size, and number of epochs), in a way that scores in the vicinity are consistent with the highest validation score but allowing for variation in order to estimate standard errors. We follow this process for 3 different random seeds for each dataset. We obtained 5154 fine-tuned models across datasets and pretrained models.

Table 1 shows a summary of all the variables

[2] Therefore, at level 2 we have 52 groups: 4 choices of pretrained models (including BERT) by 13 downstream datasets.

[3] For instance, we may expect random seeds to have a different effect, due to chance, on test scores depending on the choice of group, i.e. depending on the choice of pretrained model and dataset; thus, for each group, we can estimate the number of points, represented by a γ_{ij} coefficient, that a random seed deviates from the average effect of that random seed across all groups, represented by a β_i coefficient.

[4] Also known as Mixed Models and Hierarchical Linear Models in other fields.

[5] We follow fine-tuning guidelines from BiomedBERT and BioLinkBERT, and we use BioLinkBERT's fine-tuning code.

[6] Models which scored the highest on the validation set of each dataset.

used for the analysis.[7] Most of the variables are indicator (binary) variables which take the value of 1 whenever that variable is used by a particular instance and zero otherwise. On the other hand, the variable num_epochs takes integer values representing the number of epochs used for fine-tuning a specific model.

3.2 Multilevel Model

We instantiate Equation 1 with the variables in Table 1. As a common goal in the literature (Frank E. Harrell, 2015), we aim to find which variables have a statistically-significant effect on downstream scores across BLURB datasets.[8] We follow model-building, hypothesis-testing, and evaluation strategies from Robson and Pevalin (2016), Sommet and Morselli (2021), and Brown (2021). To fit MLMs we use the R-package *lmerTest* (Kuznetsova et al., 2017). We use the statistical tests from *lmerTest* to compute significance values ($\alpha = 0.05$ level), AIC, and BIC scores.[9] Furthermore, to estimate the proportion of explained variability in test scores by our variables we compute R-squared effects using the framework of Rights and Sterba (2019) via the R-package *r2mlm* (Shaw et al., 2022).

We added an additional term to our MLM not shown in Equation 1: interaction terms between level-2 variables; these terms are of the form $\beta_m(x_i \times x_k)$, which will help us see if a variable behaves differently for particular datasets in Section 4.

4 Multilevel Analysis and Results

We show the results of fitting our MLM. For Tables 2, 3, and 4, the statistical significance code is: p=0 '***', p<0.001 '**', p<0.01 '*'.

MLM results for level-1 variables: We first test for the statistical significance of fixed- and random-effects of level-1 variables. We observe in Table 2 that the fixed-effect of only one variable is significant, namely lr_1; this means that the learning rate of 1e-5 has a significant effect across models and datasets: models fine-tuned with this learning rate, on average, will lose 1 downstream point as shown by the coefficient of lr_1. We also see that the random seeds seed_20 and seed_47 have a small, positive impact on test scores, on average, across models and datasets; nevertheless, these fixed-effects seem to be due to chance since they are not statistically significant. However, likelihood ratio tests show that all random coefficients are statistically significant (Table 4). This means that level-1 variables behave in different ways for each group (combination of pretrained model and dataset) as we explain below.

Does chance play a role? All level-1 variables behave differently for each pretrained model; but, we note in particular that seed_20 and seed_47 contribute the biggest variability in test scores as seen in Table 4: on average, scores vary up to ($\pm$) 4.79 and ($\pm$) 6.21 points due to the choice of random seed.[10] If we average all the random coefficients[11] of seed_20 and seed_47 for each pretrained model across datasets, we find that BioBERT loses 2.33 and 3 points when using such random seeds. However, BiomedBERT and BioLinkBERT gain 0.52, 0.22 and 0.09, 0.64 points, respectively, due to such randomness.

MLM results for level-2 variables: We observe that most level-2 variables are statistically significant (Table 2), such as the effects of all datasets, meaning that different datasets lead to different results. Also significant are the contributions from BioBERT and BiomedBERT, namely, DAPT and Pretrain_PubMed, respectively, meaning that their effects –an average gain of 2.25 and 4.36 points with respect to vanilla BERT– are consistent across datasets. Surprisingly, the Link function is not significant: probably, its effect is not systematic across datasets. To better understand its effect, we estimated its interaction with all datasets; as we see in Table 3, when the Link function is used with QA datasets, its effect is remarkable: models fine-tuned with BioASQ and PubMedQA datasets gain, on avg., 8.62 and 3.68 points, respectively. However, this figure does not happen with any other dataset. Moreover, the effect of the Link function, besides non-significant, is rather small, which means that whenever the Link function is used, on average, we

[7] We include variables for the downstream datasets to take into account the fact that some datasets may be more difficult than others which may impact on the scores.

[8] We chose an MLM over simple linear regression since 1) it allows for multiple levels of analysis, and 2) the fine-tuned models inside a group are not independent from each other and only MLMs can account for such non-independence.

[9] We prefer models that decrease AIC or BIC scores.

[10] These figures represent a comparison of how much variability seed_20 and seed_47 introduce in the test scores with respect to the variability introduced by seed_59.

[11] We do not display the random coefficients since we believe it is more informative to provide an aggregated estimate.

Variable name	Level	Description
seed_20, seed_47, seed_59	1	Random seeds used for fine-tuning the Biomedical models
lr_1, lr_2, lr_3, lr_4, lr_5	1	Learning rates used for fine-tuning the Biomedical models
batch_16, batch_32	1	Batch sizes used for fine-tuning the Biomedical models
num_epochs	1	Number of epochs for fine-tuning the Biomedical models
BioBERT	2	Indicator variable for BioBERT
BiomedBERT	2	Indicator variable for BiomedBERT
BioLinkBERT	2	Indicator variable for BioLinkBERT
DAPT	2	Indicator of Domain Adaptive Pretraining on BERT
Pretrain_PubMed	2	Indicator of pretraining BERT with PubMed data from scratch
Link	2	Indicator variable of BioLinkBERT's Link function
all datasets names	2	Indicator variables of the datasets in the BLURB suite

Table 1: Variables used to model the variability in downstream scores for target Biomedical NLP models across datasets in the BLURB suite. Level 1 corresponds to fine-tuning; level 2 to pretraining; all datasets names: BC2GM, BC5_chem, BC5_disease, NCBI, JNLPBA, PICO, ChemProt, DDI, GAD, BIOSSES, HoC, BioASQ, PubMedQA.

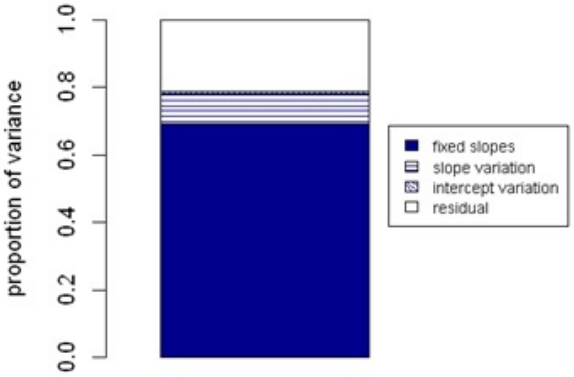

Figure 1: R-squared: Decomposition of variance across fixed and random effects.

would only see an improvement of 0.07 points on any downstream dataset.

Effects from pretrained models: If we fit our MLM with indicator variables for each pretrained model, instead of their contributions, we obtain the following effects: BioBERT (1.79**), BiomedBERT (4.52***), BioLinkBERT (3.47***). The result for BioLinkBERT seems to contradict the non-significant effect of the Link function. It does not. The effect of the BioLinkBERT variable takes into account all possible functions inside BioLinkBERT (without disentangling them) including the variable of Pretrain_PubMed since BioLinkBERT was pretrained from scratch with PubMed data. This means that, overall, BioLinkBERT is highly useful: it surpasses vanilla BERT, on average, by 3.47 points across datasets, though the Link function does not seem to be the main reason for this result due to its small effect size and lack of statistical significance. Surprisingly, we see that BiomedBERT has the biggest mean effect of all models: 4.52 points improvement over BERT.

R-squared effects: As shown in Figure 1, fixed-effects of level-1 and level-2 variables account for around 70% of all the variability in the test scores; however, given that most of the level-1 coefficients are non-significant and moderately small, we would expect them to contribute little to this explanation of variability. Surprisingly, though, level-1 random coefficients (slope variation) account for around 10% of the variance in test scores, a considerable portion of the variability. Finally, we note that around 20% of the variance remains unexplained (the residual part) which may mean two things. First, there is still room for adding variables at either level to better explain the test scores; we hypothesized that other pretraining features, such as batch size, could impact on the scores, however, it was not possible to add them to the analysis since they perfectly correlate with variables already added, leading to the problem of collinearity. And second, fully understanding NLP models is a complex task which requires detailed analyses of several variables.

Robust estimates: As shown in Table 2, most of the standard errors (SEs) are small which means that our coefficients estimates are robust, i.e. their estimation is precise due to the low variability represented by the corresponding SE, something that could be more difficult to achieve when only averaging scores from a handful of models across random seeds as is usual in the NLP literature.

Variable	Coeff. (β)	SE	t
Intercept	53.96***	0.88	61.09
seed_20	0.49	0.70	0.70
seed_47	0.21	0.89	0.24
lr_1	-1.00**	0.35	-2.82
lr_2	0.59	0.51	1.14
lr_3	0.65	0.48	1.35
lr_4	0.15	0.24	0.61
batch_16	0.28	0.17	1.59
num_epochs	-0.02	0.02	-1.00
DAPT	2.25***	0.38	5.90
Pretrain_PubMed	4.36***	0.41	10.41
Link	0.07	0.32	0.21
BC2GM	27.40***	2.02	13.54
BC5_chem	35.42***	0.60	58.49
BC5_disease	25.96***	0.64	40.21
NCBI	31.79***	0.73	43.29
JNLPBA	21.26***	0.82	25.92
PICO	16.14***	0.89	18.00
ChemProt	17.56***	0.73	23.97
DDI	23.07***	0.81	28.23
GAD	23.81***	0.57	41.45
BIOSSES	20.85***	0.67	30.74
HoC	25.48***	0.75	33.59
BioASQ	17.58***	0.87	20.16

Table 2: Results of MLM: fixed-effects of variables at levels 1 and 2. Coeff: coefficient. SE: Standard Error. t: t-value. We use seed_59, lr_5, batch_32, and PubMedQA as baselines to avoid collinearity.

Interaction	Coeff. (β)	SE	t
DAPT×BIOSSES	-7.49***	0.71	-10.5
PubMed×NCBI	-2.31*	0.86	-2.6
PubMed×PICO	-2.22*	0.95	-2.3
PubMed×BIOSSES	13.1***	0.63	20.5
PubMed×BioASQ	7.37***	1.49	4.9
Link×HoC	-3.58***	0.85	-4.2
Link×BioASQ	8.62***	1.83	4.6
Link×PubMedQA	3.68*	1.38	2.65

Table 3: Results of MLM: interaction terms. PubMed stands for Pretrain_PubMed. Only statistically significant interactions are displayed.

Variable	Variance	Std. Dev.
Intercepts	9.29***	3.04
seed_20	22.96***	4.79
seed_47	38.59***	6.21
lr_1	3.03***	1.74
lr_2	9.72***	3.11
lr_3	8.78***	2.96
lr_4	0.42*	0.64
batch_16	0.37**	0.60
num_epochs	0.01**	0.10

Table 4: Results of MLM: random effects (random intercepts and random coefficients). Variables seed_59, lr_5, batch_32 are used as baselines to avoid collinearity.

5 Conclusions

Our multilevel analysis of Biomedical models can disentangle the effects from fine-tuning and pretraining by providing particular effects of each variable with respective statistical significance. As we saw, contrary to expectation, BiomedBERT has the biggest mean effect across datasets from all models. Moreover, even though BioLinkBERT holds as a useful model, its main contribution –the Link function– does not seem to be the main reason for its success, except for QA datasets where the Link function excels. Furthermore, we showed that all fine-tuning variables behave differently for each pretrained model, giving some advantage to some models purely by chance. And this figure, according to R-squared tests, accounts for 10% of all the test scores; thus, we suggest using several random seeds to counterbalance their effects. Finally, we note that it would be nearly impossible to see all these figures with current evaluation methods –a single mean score. We hope the community will adopt MLMs as a deeper evaluation method.

Limitations

We note that there may be more independent variables having an effect on downstream scores that we did not take into account due to their difficulty to be measured, or to be known, such as detailed pretraining hyperparameters or data pre-processing methods. Also, we note that our design of the problem as a 2-level hierarchy may not be the most optimal design; there are more design types that can be operationalized via MLMs; however, hierarchical models are the most common and studied in the literature. Furthermore, in this paper, we fine-tuned the base sizes of the pretrained models (e.g. BioLinkBERT-base), we did not analyze the large-size models (e.g. BioLinkBERT-large) which we leave as future work. Also, due to GPU memory limitations, we did not explore more levels of the variables studied such as using a batch of size 64 for fine-tuning which may benefit some of the models.

References

R.H. Baayen, D.J. Davidson, and D.M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412. Special Issue: Emerging Data Analysis.

Violet A. Brown. 2021. An introduction to linear mixed-effects modeling in R. *Advances in Methods and Practices in Psychological Science*, 4(1):1–19.

Jan de Leeuw and Erik Meijer. 2008. *Handbook of Multilevel Analysis*, first edition. Springer New York, NY.

Jr. Frank E. Harrell. 2015. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*, second edition. Springer Cham.

Harvey Goldstein, Simon Burgess, and Brendon McConnell. 2007. Modelling the effect of pupil mobility on school differences in educational achievement. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 170(4):941–954.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

Charles M. Judd, Jacob Westfall, and David A. Kenny. 2017. Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology*, 68(1):601–625. PMID: 27687116.

Reinhold Kliegl, Michael E. J. Masson, and Eike M. Richter. 2010. A linear mixed model analysis of masked repetition priming. *Visual Cognition*, 18(5):655–681.

Reinhold Kliegl, Ping Wei, Michael Dambacher, Ming Yan, and Xiaolin Zhou. 2011. Experimental effects and individual differences in linear mixed models: Estimating the relationship between spatial, object, and attraction effects in visual attention. *Frontiers in Psychology*, 1.

Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13):1–26.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Melis Muradoglu, Joseph R. Cimpian, and Andrei Cimpian. 2023. Mixed-effects models for cognitive development researchers. *Journal of Cognition and Development*, 24(3):307–340.

Hugo Quené. 2008. Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo. *The Journal of the Acoustical Society of America*, 123(2):1104–1113.

Jon Rasbash, George Leckie, Rebecca Pillinger, and Jennifer Jenkins. 2010. Children's Educational Progress: Partitioning Family, School and Area Effects. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 173(3):657–682.

J. D. Rights and S. K. Sterba. 2019. Quantifying explained variance in multilevel models: An integrative framework for defining r-squared measures. *Psychological Methods*, 24(3):309–338.

Karen Robson and David Pevalin. 2016. *Multilevel Modeling in Plain Language*, first edition. SAGE Publications Ltd.

Mairead Shaw, Jason D. Rights, Sonya S. Sterba, and Jessica Kay Flake. 2022. r2mlm: An r package calculating r-squared measures for multilevel models. *Behavior Research Methods*, 55:1942–1964.

Nicolas Sommet and Davide Morselli. 2021. Keep calm and learn multilevel linear modeling: A three-step procedure using spss, stata, r, and mplus. *International Review of Social Psychology*, 34(1).

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. LinkBERT: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8003–8016, Dublin, Ireland. Association for Computational Linguistics.