RESEARCH ARTICLE

Ablations over transformer models for biomedical relationship extraction [version 1; peer review: 2 approved with reservations]

Richard G Jackson , Erik Jansson, Aron Lagerberg, Elliot Ford , Vladimir Poroshin, Timothy Scrivener, Mats Axelsson, Martin Johansson, Lesly Arun Franco , Eliseo Papa

AstraZeneca PLC, Eastbrook House, Cambridge, CB2 8DU, UK

V1 First published: 16 Jul 2020, 9:710
<https://doi.org/10.12688/f1000research.24552.1>
Latest published: 16 Jul 2020, 9:710
<https://doi.org/10.12688/f1000research.24552.1>

Abstract

Background: Masked language modelling approaches have enjoyed success in improving benchmark performance across many general and biomedical domain natural language processing tasks, including biomedical relationship extraction (RE). However, the recent surge in both the number of novel architectures and the volume of training data they utilise may lead us to question whether domain specific pretrained models are necessary. Additionally, recent work has proposed novel classification heads for RE tasks, further improving performance. Here, we perform ablations over several pretrained models and classification heads to try to untangle the perceived benefits of each.

Methods: We use a range of string preprocessing strategies, combined with Bidirectional Encoder Representations from Transformers (BERT), BioBERT and RoBERTa architectures to perform ablations over three RE datasets pertaining to drug-drug and chemical protein interactions, and general domain relationship extraction. We explore the use of the RBERT classification head, compared to a simple linear classification layer across all architectures and datasets.

Results: We observe a moderate performance benefit in using the BioBERT pretrained model over the BERT base based model, although there appears to be little difference when comparing BioBERT to RoBERTa large. In addition, we observe a substantial benefit of using the RBERT head on the general domain RE dataset, but this is not consistently reflected in the biomedical RE datasets. Finally, we discover that randomising the token order of training data does not result in catastrophic performance degradation in our selected tasks.

Conclusions: We find a recent general domain pretrained model performs approximately the same as a biomedical specific one, suggesting that domain specific models may be of limited use given

Open Peer Review

Approval Status ? ?

	1	2
version 1		
16 Jul 2020	view	view

1. **Yuan Li**, The University of Melbourne, Melbourne, Australia

Karin M. Verspoor , The University of Melbourne, Melbourne, Australia
RMIT University, Melbourne, Australia

2. **Jens Dörpinghaus** , German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany

Any reports and responses or comments on the article can be found at the end of the article.

the tendency of recent model pretraining regimes to incorporate ever broader sets of data. In addition, we suggest that care must be taken in RE model training, to prevent fitting to non-syntactic features of datasets.

Keywords

Natural Language Processing, Biomedical Relationship Extraction, NLP, ChemProt, Drug Drug Interactions, Semeval 2010 Task 8

Corresponding author: Richard G Jackson (richg.jackson@gmail.com)

Author roles: **Jackson RG:** Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Jansson E:** Conceptualization, Formal Analysis, Investigation, Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Lagerberg A:** Conceptualization, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Ford E:** Writing – Review & Editing; **Poroshin V:** Writing – Review & Editing; **Scrivener T:** Writing – Review & Editing; **Axelsson M:** Writing – Review & Editing; **Johansson M:** Writing – Review & Editing; **Franco LA:** Writing – Review & Editing; **Papa E:** Conceptualization, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: All authors are employed by AstraZeneca PLC.

Grant information: This work was funded by AstraZeneca PLC.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2020 Jackson RG *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Jackson RG, Jansson E, Lagerberg A *et al.* **Ablations over transformer models for biomedical relationship extraction [version 1; peer review: 2 approved with reservations]** F1000Research 2020, 9:710 <https://doi.org/10.12688/f1000research.24552.1>

First published: 16 Jul 2020, 9:710 <https://doi.org/10.12688/f1000research.24552.1>

Introduction

The biomedical literature is a vast corpus of unstructured facts and findings, which need to be synthesised in some systematic way in order for drug discovery scientists to make informed, logical choices about what directions and experiments to pursue. A highly valued goal of biomedical natural language processing (NLP) is to perform relationship extraction (RE) between entities of interest¹, such that the knowledge entombed within the literature can be exploited by technological solutions, such as knowledgebase representations. In recent years, groups such as BioCreative and SemEval have coalesced the community around shared RE tasks, in order that we might benchmark our methods against common standards.

Since the early forays into transfer learning to the advent of transformer based models^{2,3}, language modelling and, more recently, masked language modelling is the de rigour methodology in current NLP research. From investigations into the optimal learning objective, to explorations into the limit of pre-training, to permuting the classification head, a bewildering array of research has rapidly emerged, concerning almost every aspect of language modelling. This has created a vast experimental space for the community to explore how such developments relate to biomedical NLP.

The seminal masked language model, Bidirectional Encoder Representations from Transformers (BERT)⁴, helped to popularise the idea of pretraining on general linguistic data and subsequently fine tuning to tailor the model to downstream tasks. Pretraining is the task of learning some representation of language, such that a piece of text can be encoded into high dimensional space, representing some knowledge about how the tokens within such text relate to each other. Offshoots of BERT, such as SciBERT⁵, BioBERT⁶ and BlueBERT⁷, demonstrated that pretraining on scientific literature allow for better representations of the scientific sublanguage, leading to performance increases in downstream tasks pertaining to that domain. Work such as RoBERTa⁸ and T5⁹ further recognised that BERT had been undertrained and built upon the original architecture with an expanded pretraining procedure and a larger parameter space.

Although performance gains from larger models and lengthier pretraining is an interesting phenomenon, this represents practical issues for those working within niche domains who desire models pretrained on specific styles of document. With the rapid evolution of new architectures, and substantial costs involved in pretraining, the investment in performing domain specific pretraining becomes hard to justify when the end result may be obsolete within months. Thus, it is desirable to know whether the performance gains from domain specific pretraining outlive the original model architecture (compared to newer architectures that do not benefit from learning better representations of a domain, but perhaps benefit from learning better representations of domain independent, fundamental aspects of language).

A second aspect of language modelling concerns how model are fine-tuned to perform certain tasks. For instance, sentence

classification tasks with the original BERT model is possible by passing the sentence representation token (denoted [CLS]) through a linear layer. More recent work (specific to the task of relationship extraction) has explored how combining embedded entity information with such sentence representations can lead to significant performance boosts (the RBERT head)¹⁰. However, evidence has since emerged¹¹ that at least some of the perceived performance gains of transformer style models is due to so-called ‘Clever Hans’ type effects, where the model is fine-tuned to learn unintended correlations in datasets rather than a generalised representation of the task. This in turn raises questions about the validity of such approaches in the task of relationship extraction, and how to manufacture appropriate datasets.

The goal of this article is to attempt to address some of these questions via ablation studies of a range of popular masked language models and classification heads, to determine their performance on the task of biological relationship extraction.

Methods

We experiment with the general purpose pretrained BERT, model, the biomedical domain specific pretrained model, BioBERT, and the more recent general purpose RoBERTa model. Both BioBERT and RoBERTa are particularly relevant to the ablation tests in this study and serve as an example of a domain-specific model, and a larger model that has undergone lengthier pretraining, respectively. We combine these pretrained models with two classification heads; the commonly used linear layer based on the sentence vector produced by the final layer, and the RBERT classification head. In addition, we examine the effect of four string preprocessing techniques (two per classification head), to investigate how the differing transformer architectures respond to ablations.

Datasets

We consider ablations over three different corpora labelled with named entities and relationships. The ChemProt¹² dataset was originally created for the BioCreative VI workshop, and sought to challenge teams to deliver systems that extracted chemical protein relationships from the scientific literature. It consists of a set of 15,739 relationships annotations from 1,682 PubMed abstracts, divided into training, development and evaluation sets. The dataset covers 11 different label types, although only five undirected relationship types are used in the official evaluation. An official evaluation script is provided.

The DDI (Drug/Drug Interaction) corpus (hereafter DDI) was created for the SemEval-2013 DDI Extraction 2013 challenge¹³, and seeks to provide a dataset to support the development of NLP systems to extract various types of drug/drug interaction. It consists of 5,028 sentence level relationships manually annotated from Medline and DrugBank, labelled with one of five undirected classes (four describing different types of interaction and one null relationship class) and split into training and evaluation sets. The distribution of labels in this corpus is heavily weighted towards null relations, and thus the imbalance of classes represents an interesting problem for

ML classifiers in its own right. The provided official evaluation script calculates the macro F1 over four relationship classes (the null relation is not considered).

Finally, we make use of the Semeval 2010 Task 8 corpus¹⁴ (hereafter Semeval), which is a general English RE dataset collected from the Web, and uses a more abstract relationship classification schema than ChemProt or DDI. Here, ten relationship classes are used to annotate 10,717 sentences, which are split into a training set of 8,000 and an evaluation set of 2,717. The provided official evaluation script calculates the macro F1 over nine of the classes in a bi-directional fashion for a total of 18 classes.

For consistency with existing literature, we report our official scores (using official evaluation scripts provided with each dataset) but focus our analysis around cross validation on each set, in order to assess the consistency of the corpora and the effect of random seeds. Here, we report the mean macro averaged F1 score with a five-fold cross validation split.

Pretrained model selection

Originally, we planned to conduct an analysis comparing a wide range of transformer architectures. However, our preliminary investigations suggested that many were too cumbersome to work with, either in terms of compute required, the quality of the pretrained model or the maturity of the codebase. To this end, we restricted our analysis to the pretrained models BERT Base, BioBERT 1.1, RoBERTa base and RoBERTa large, as described in Table 1. Our principal question compares the evaluation performance of BERT Base, BioBERT, and RoBERTa base, as models of approximately equal parameter counts. However, we additionally decided to include RoBERTa large to explore any potential benefits from using a larger model with a higher quality pretraining regime (based upon General Language Understanding Evaluation benchmark results¹⁵).

All experiments were conducted with the HuggingFace Transformers implementations, version 2.4.1¹⁶

Table 1. Pretrained models parameter count and training corpora.

Model name	Version	Parameter count	Training corpora
BERT	base	110m	BooksCorpus English Wikipedia
BioBERT	1.1	110m	As BERT with PubMed abstracts
RoBERTa	base	120m	BookCorpus English Wikipedia CC-News Open Web Text Stories
RoBERTa	large	355m	same as RoBERTa base

Classification heads

Pretrained models are frequently employed in classification tasks, wherein a linear layer is constructed on top of the final layer. Recently, some modifications of this approach have been proposed, to combine specific entity information into the classification layer, to support relationship classification tasks. Wu and He¹⁰ suggested averaging the token pieces representing each entity, and concatenating the output with the sentence vector before applying a fully connected feed forward layer, giving rise to the RBERT classification head and setting a new benchmark in the Semeval 2010 Task 8 dataset. In this work, we compare both the simple linear layer classification head and the RBERT head.

Preprocessing

RE is commonly construed as a sentence classification task, wherein the label assigned to the relationship between two entities in a sentence are instead assigned to the sentence. However, such an approach can be problematic; for instance, if there are more than two entities in a sentence, and/or more than two relationships (a common occurrence in biomedical text), leading to a situation where the same sentence can yield two conflicting labels.

To mitigate this, various strategies have been used, such as substituting the entities of interest with nominal placeholder tokens, such that all strings seen by a classifier are unique and creating the possibility for a classifier to learn the syntactic importance of the placeholder tokens with regard to the relationship that binds them¹⁷. In contrast, the RBERT architecture depends on inserting special characters around the two entities of interest, to inform the classifier of the two input entities without removing information about the entity itself.

Here, we employ ablations on these preprocessing strategies depending on the type of classification head used with the pretrained model (Table 2).

The purpose of the sentence splitting ablation is to provide a baseline classification performance for the underlying pretrained model, without any special characteristics applied to the entities of interest (note, all other transformations include this sentence splitting step). The placeholder transformation is a commonly used strategy in RE^{6,18,19}, where the entities in question are masked by some arbitrary token, thereby attempting to reduce overfitting of the classifier and allowing different relationships between different entity pairs in the same sentence to be represented. Similarly, the bounding special characters ablation is the original transformation as described in the original RBERT paper, whereas the purpose of the masked bounding special character transformation is to remove any entity specific information from the RBERT head. By removing this entity information, our intent is to explore the extent to which the positional information of the entity pairs are used in making the relationship classification, as opposed to the entity embedding information of the entity pairs.

Since some of the preprocessing strategies can lead to undesirable mutations of the underlying data (for instance, it is not possible to represent discontinuous entity boundaries, or overlapping entity boundaries for the placeholder or bounding special character strategies), we filter out any such instances that cannot be transformed for all pretrained model/classification head configurations, such that our training and evaluation sets are consistent across all experiments.

Training

In this ablation study, we aspire for consistency across experiments, rather than attempting to optimise for overall evaluation performance across our selected datasets. To this end, we do not attempt a hyper parameter search. Instead, we defer to recommended hyperparameters for classification tasks based upon the General Language Understanding Evaluation benchmark, as described in the original BERT and RoBERTa papers (Table 3).

One important consideration in hyperparameter selection is the maximum sequence length used. Naturally, it is desirable to use a sequence length big enough to enable the longest

sentence in each dataset to be passed through the model. However, longer sequence lengths rapidly increase the memory usage in GPUs, and thus a variable batch size must be selected as required for a given dataset. Since larger batch sizes tend to be desirable²⁰, we originally sought to specify a minimum batch size of 16 across all experiments, in line with recommendations in the BERT and ROBERTA papers. However, initial experiments uncovered that larger models such as RoBERTa large were unable to handle the required sequence length and batch size on the hardware available to us (Tesla V100 16Gb GPUs). To overcome this, we reduced the batch size to four and used eight gradient accumulation steps in all experiments.

We executed six runs across each dataset, per experiment configuration. The first run used the official train/test splits as described in the original datasets, whereas the remaining five runs were comprised of cross validation runs, varying the random seed between folds.

We trained for a maximum of five epochs, and after the first epoch, implemented an early stopping regime that tested for improvements in the average micro F1 score across all classes,

Table 2. String transformation ablations.

String preprocessing method	Applies to	Description	Example
sentence splitting	Linear layer classification head	No modification of the original text	The cat sat on the mat
placeholder substitution	Linear layer classification head	For each entity pair, the entities are replaced with unique strings	The ent1 sat on the ent2
bounding special characters	RoBERT head	For each entity pair, the entities are surrounded with special characters	The \$ cat \$ sat on the # mat #
masked bounding special characters	RoBERT head	For each entity pair, the entities are surrounded with special characters and the entity token(s) themselves are replaced with unique strings	The \$ ent1 \$ sat on the # ent2 #

Table 3. Model hyperparameters used in training.

Hyperparameter	BERT models	RoBERTa models
Batch size	4	4
Gradient accumulation steps	8	8
Adam Epsilon	1e-6	1e-6
Adam Beta 1	0.9	0.9
Adam Beta 2	0.999	0.980
Weight decay	0.1	0.1
Learning rate	5e-5	2e-5
Max training epochs	5	5
GPU count	4 * V100 16Gb	4 * V100 16Gb

after every 5% of the dataset. Five successive failures to improve the F1 resulted in the termination of training, and we logged the highest macro F1 scores reached during training for our cross validation results.

Results and discussion

The results of each of our ablations are presented in [Figure 1](#) (tabularised in [Table 4](#)).

Effect of baseline model

With respect to the differences with BERT base cased and BioBERT, we observe a moderate benefit by using the BioBERT model on the biomedical Chemprot and DDI datasets, and a moderate benefit by using BERT base on SemEval, in line with observations that domain specific training can improve performance. However, BioBERT and RoBERTa large appear to be approximately equivalent across all datasets, with

Figure 1. Cross validation results. BERT_BC = BERT base cased, BERT_BIO = bioBERT, ROBERTA_B = RoBERTa base, ROBERTA_L = RoBERTa large, PH = placeholder, SSplit = sentence splitter, SpChar = bounding special characters, MSpChar = masked bounding special characters.

Table 4. Full results across experiments. BERT_BC = BERT base cased, BERT_BIO = bioBERT, ROBERTA_B = RoBERTa base, ROBERTA_L = RoBERTa large, PH = placeholder, SSplit = sentence splitter, SpChar = bounding special characters, MSpChar = masked bounding special characters. F1 values represents macro F1. Std = standard deviation for cross validation. Random token results use randomly ordered tokens in training data (evaluation data is kept intact).

Model type	Dataset	String transformation	x-val f1 mean	x-val f1 std	Best official result	Random token x-val mean	Random token x-val std
BERT_BC	chemprot	PH	0.817	0.025	0.769	0.653	0.018
BERT_BC	chemprot	SSplit	0.684	0.018	0.655	0.653	0.028
BIO	chemprot	PH	0.882	0.010	0.781	0.676	0.020
BIO	chemprot	SSplit	0.716	0.015	0.687	0.663	0.016
RoBERTa_B	chemprot	PH	0.819	0.018	0.748	0.633	0.052
RoBERTa_B	chemprot	SSplit	0.678	0.023	0.662	0.601	0.033
RoBERTa_L	chemprot	PH	0.888	0.013	0.795	0.665	0.015

Model type	Dataset	String transformation	x-val f1 mean	x-val f1 std	Best official result	Random token x-val mean	Random token x-val std
RoBERTa_L	chemprot	SSplit	0.689	0.015	0.669	0.631	0.029
BERT_BC	chemprot	SpChar	0.846	0.010	0.736	~	~
BERT_BC	chemprot	MSpChar	0.846	0.005	0.748	~	~
BIO	chemprot	SpChar	0.886	0.004	0.777	~	~
BIO	chemprot	MSpChar	0.875	0.009	0.775	~	~
RoBERTa_B	chemprot	SpChar	0.333	0.067	0.427	~	~
RoBERTa_B	chemprot	MSpChar	0.283	0.063	0.412	~	~
RoBERTa_L	chemprot	SpChar	0.871	0.016	0.791	~	~
RoBERTa_L	chemprot	MSpChar	0.859	0.023	0.77	~	~
BERT_BC	ddi	PH	0.777	0.039	0.7327	0.485	0.032
BERT_BC	ddi	SSplit	0.439	0.025	0.2793	0.412	0.046
BIO	ddi	PH	0.818	0.050	0.757	0.507	0.011
BIO	ddi	SSplit	0.459	0.020	0.3499	0.426	0.024
RoBERTa_B	ddi	PH	0.745	0.067	0.6238	0.343	0.062
RoBERTa_B	ddi	SSplit	0.430	0.025	0.3155	0.318	0.079
RoBERTa_L	ddi	PH	0.804	0.056	0.7648	0.371	0.020
RoBERTa_L	ddi	SSplit	0.441	0.026	0.4049	0.406	0.029
BERT_BC	ddi	SpChar	0.793	0.041	0.749	~	~
BERT_BC	ddi	MSpChar	0.802	0.041	0.756	~	~
BIO	ddi	SpChar	0.837	0.044	0.7692	~	~
BIO	ddi	MSpChar	0.828	0.043	0.7803	~	~
RoBERTa_B	ddi	SpChar	0.334	0.026	0.1609	~	~
RoBERTa_B	ddi	MSpChar	0.327	0.074	0.1245	~	~
RoBERTa_L	ddi	SpChar	0.828	0.037	0.7674	~	~
RoBERTa_L	ddi	MSpChar	0.816	0.027	0.7781	~	~
BERT_BC	semeval	PH	0.682	0.033	0.7926	0.523	0.032
BERT_BC	semeval	SSplit	0.671	0.016	0.783	0.551	0.080
BIO	semeval	PH	0.650	0.021	0.7626	0.457	0.051
BIO	semeval	SSplit	0.653	0.035	0.7575	0.520	0.012
RoBERTa_B	semeval	PH	0.648	0.011	0.786	0.333	0.077
RoBERTa_B	semeval	SSplit	0.601	0.018	0.7263	0.391	0.057
RoBERTa_L	semeval	PH	0.710	0.016	0.8047	0.415	0.085
RoBERTa_L	semeval	SSplit	0.692	0.033	0.7744	0.488	0.032
BERT_BC	semeval	SpChar	0.827	0.022	0.8752	~	~
BERT_BC	semeval	MSpChar	0.716	0.038	0.8138	~	~
BIO	semeval	SpChar	0.797	0.025	0.8557	~	~
BIO	semeval	MSpChar	0.714	0.017	0.8012	~	~
RoBERTa_B	semeval	SpChar	0.240	0.029	0.3663	~	~
RoBERTa_B	semeval	MSpChar	0.198	0.013	0.312	~	~
RoBERTa_L	semeval	SpChar	0.838	0.027	0.8878	~	~
RoBERTa_L	semeval	MSpChar	0.717	0.050	0.8297	~	~

RoBERTa large ranking marginally higher in most experiments. The surprisingly poor performance of the RoBERTa base model compared to BERT base suggests that most of RoBERTa large's performance is due to the higher parameter count, rather than the larger size of RoBERTa's pretraining corpora. Nevertheless, given the very poor performance of the RoBERTa base model with the RBERT head, we are unable to rule out other factors. Particularly difficult to separate is the benefit of training on domain specific data; although RoBERTa is not a biomedical specific model, we examined the contents of the OpenWebText corpora upon which it is trained, and discovered over 11,000 references to PubMed abstracts, as well as other references to providers of scientific literature, suggesting that some of RoBERTa's performance on biomedical text may come from partial exposure to the domain during pretraining.

Effect of classification head and string transformation

On the biomedical datasets, the RBERT classification head seems to provide a small benefit in the DDI task, but this is not observed on the classification performance of ChemProt compared to the placeholder string transformation with the linear layer classifier head. However, the RBERT head appears to substantially boost performance on the SemEval dataset, although the benefits are substantially reduced if entity information is masked. In the case of SemEval dataset, this suggests that the classifier is making more use of the contextual entity embedding than the positional information of the token, and is therefore reliant on latent correlations between the entity pairs and the label, rather than an interpretation of the syntax

of the sentence. In the case of the biomedical datasets, many of the classification head/string transformations performed similarly, suggesting none of these are particularly important and that the attention mechanism itself is mostly responsible for learning a representation of the data. A potentially related finding from our results is that even simple sentence classifiers give reasonable performance on the ChemProt and SemEval datasets, with no knowledge of what entity pairs in a sentence the label refers to. To explore this further, we randomised the token order for each instance in the training sets and repeated our experiments for the sentence splitter and placeholder string transformations (Figure 2). Although this ablation created a marked drop in performance across all datasets, we were surprised that this drop was not as substantial as we might have expected. By removing all syntactic information from the training data, it would appear (to a varying degree) that the classifiers are still able to learn some aspects of the relationship classification task using only contextualised embedding information.

We suspect that this effect is likely to be attributed to the nature of the underlying training data. Although the attention mechanism employed by the models we tested should be able to learn the required syntactic relationships in order to perform the RE task²¹, it is also possible for them to learn other aspects of the training data that correlate the sentence embedding information with the given label. For instance, it seems likely that the the presence of certain words may occur more frequently with certain label types, such as verbs suggesting

Figure 2. Cross validation results with tokens randomly ordered in training data. BERT_BC = BERT base cased, BERT_BIO = bioBERT, ROBERTA_B = RoBERTa base, ROBERTA_L = RoBERTa large, PH = placeholder, SSplit = sentence splitter. The horizontal blue line indicates the expected performance of a random classifier.

gene regulation activities in the case of ChemProt. Such an effect has recently been established for various NLP architecture across natural language inference datasets, including BERT²². Therefore, models that are trained to make use of such non-syntactic information probably generalise poorly, although further work will be required to establish this conclusively.

Conclusions

In this study, we perform a variety of ablations over an array of models and configurations over three RE datasets. We find that there are benefits in using models pretrained on biomedical text, but the benefits tend to be relatively small and/or task specific on the datasets we explored. Further, there is a tendency of newer models to be trained on larger corpora of text, which appear to encompass the biomedical domain. Future work might revisit analyses such as ours, to determine whether the benefits of domain specific model training outweigh the costs. Finally, we suggest that care must be taken in the training of models for RE, as it appears likely that classifiers are susceptible to overfitting on non-syntactic features. This may be alleviated by the creation of training data that depend heavily on syntactic features, and advancing

other methodologies such as data augmentation and Universal Adversarial Triggers²³.

Data availability

Source data

The DDI dataset is available from <https://github.com/isequra/DDICorpus>.

The ChemProt dataset is available from <https://biocreative.bioinformatics.udel.edu/news/corpora/chemprot-corpus-biocreative-vi/>.

The Semeval 2010 task 8 dataset is available from <http://semeval2.fbk.eu/semeval2.php?location=data>.

Code availability

Source code available from: <https://github.com/RichJackson/pytorch-transformers>

Archived source code at time of publication: <http://www.doi.org/10.5281/zenodo.3894625>²⁴

License: [Apache 2.0](#)

References

- Huang CC, Lu Z: **Community challenges in biomedical text mining over 10 years: success, failure and the future**. *Brief bioinform.* 2016; 17(1): 132–44. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Malte A, Ratadiya P: **Evolution of transfer learning in natural language processing**. *ArXiv.* abs/1910.07370. 2019. [Reference Source](#)
- Vaswani A, Shazeer N, Parmar N, et al.: **Attention is all you need**. *NIPS*. 2017. [Reference Source](#)
- Devlin J, Chang MW, Lee K, et al.: **Bert: Pre-training of deep bidirectional transformers for language understanding**. *NAACL-HLT*. 2019. [Publisher Full Text](#)
- Beltagy I, Cohan A, Lo K: **Scibert: Pre-trained contextualized embeddings for scientific text**. *ArXiv.* abs/1903.10676, 2019. [Reference Source](#)
- Lee J, Yoon W, Kim S, et al.: **Biobert: a pre-trained biomedical language representation model for biomedical text mining**. *Bioinformatics*. 2019. [PubMed Abstract](#) | [Publisher Full Text](#)
- Peng Y, Yan S, Lu Z: **Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets**. *BioNLP@ACL*. 2019. [Publisher Full Text](#)
- Liu Y, Ott M, Goyal N, et al.: **Roberta: A robustly optimized bert pretraining approach**. *ArXiv.* abs/1907.11692. 2019. [Reference Source](#)
- Raffel C, Shazeer N, Roberts A, et al.: **Exploring the limits of transfer learning with a unified text-to-text transformer**. *ArXiv.* abs/1910.10683. 2019. [Reference Source](#)
- Wu S, He Y: **Enriching pre-trained language model with entity information for relation classification**. *CIKM '19*. 2019; 2361–2364. [Publisher Full Text](#)
- Niven T, Kao HY: **Probing neural network comprehension of natural language arguments**. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019; 4658–4664. Florence, Italy, Association for Computational Linguistics. [Publisher Full Text](#)
- Krallinger M, Rabal SA, et al.: **Overview of the biocreative vi chemical-protein interaction track**. 2017. [Reference Source](#)
- Herrero-Zazo M, Segura-Bedmar I, Martínez P, et al.: **The ddi corpus: An annotated corpus with pharmacological substances and drug-drug interactions**. *J Biomed Inform.* 2013; 46(5): 914–20. [PubMed Abstract](#) | [Publisher Full Text](#)
- Hendrickx I, Kim SN, Kozareva Z, et al.: **Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals**. *ArXiv.* 2009. abs/1911.10422. [Reference Source](#)
- Wang A, Singh A, Michael J, et al.: **Glue: A multi-task benchmark and analysis platform for natural language understanding**. *BlackboxNLP@EMNLP*. 2018. [Reference Source](#)
- Wolf T, Debut L, Sanh V, et al.: **Hug-gingface's transformers: State-of-the-art natural language processing**. *ArXiv.* 2019. abs/1910.03771. [Reference Source](#)
- Lim S, Kang J: **Chemical-gene relation extraction using recursive neural network**. *Database (Oxford)*. 2018; 2018: bay060. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Dligach D, Miller T, Lin C, et al.: **Neural temporal relation extraction**. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. 2017; 746–751. Valencia, Spain, Association for Computational Linguistics. [Reference Source](#)
- Shi P, Lin J: **Simple bert models for relation extraction and semantic role labeling**. 2019. [Reference Source](#)
- Smith LN: **A disciplined approach to neural network hyper-parameters: Part 1 - learning rate, batch size, momentum, and weight decay**. *ArXiv.* 2018. abs/1803.09820. [Reference Source](#)
- Tenney I, Das D, Pavlick E: **Bert rediscovers the classical nlp pipeline**. *ACL*. 2019. [Publisher Full Text](#)
- Thomas McCoy R, Pavlick E, Linzen T: **Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference**. *ACL*. 2019. [Reference Source](#)
- Wallace E, Feng S, Kandpal N, et al.: **Universal adversarial triggers for attacking and analyzing nlp**. *EMNLP/IJCNLP*. 2019. [Publisher Full Text](#)
- Wolf T, Debut L, SANH V, et al.: **RichJackson/pytorch-transformers: supporting ablation paper v3**. 2020. <http://www.doi.org/10.5281/zenodo.3894625>

Open Peer Review

Current Peer Review Status: ? ?

Version 1

Reviewer Report 22 September 2021

<https://doi.org/10.5256/f1000research.27086.r93667>

© 2021 Dörpinghaus J. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Jens Dörpinghaus

German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany

This paper is mostly dedicated to NLP and RE tasks within the biomedical field. Its main approach is to compare different pretrained models BERT, BioBERT and RoBERTa large/base in particular on drug-drug, chemical protein interactions and a general dataset. The authors conclude that there is not a great difference between these approaches, and suggest preferring pretrained models to domain specific models.

The paper is well-structured and sound. The authors provide not only references to their datasets, but also to their code. The contribution of this paper is limited to the comparison of different existing methods. This is my biggest concern: Since there is no section dedicated to the state of the art or related work, the decision on comparison is somehow unsubstantiated. I encourage the authors to provide this missing section and to provide more discussion on why they chose these models and what makes them particularly interesting. Some missing references as a starting point for further review include: Hoyt *et al.* (2021)¹, Langnickel & Fluck (2021)², Madan *et al.* (2017)³, and Yadav *et al.* (2020)⁴.

The conclusions are not totally convincing due to the fact that the method and analysis are not introduced well. Why is the analysis limited to these approaches? And why do they allow such a generalization? I doubt that the results allow suggesting a preference for pretrained models to domain specific models. Maybe the authors can provide more details on that.

References

1. Hoyt C, Domingo-Fernández D, Hofmann-Apitius M: BEL Commons: an environment for exploration and analysis of networks encoded in Biological Expression Language. *Database*. 2018; 2018. [Publisher Full Text](#)
2. Langnickel L, Fluck J: We are not ready yet: limitations of transfer learning for Disease Named Entity Recognition. *bioRxiv*. 2021. [Publisher Full Text](#)
3. Madan S, Szostak J, Dörpinghaus J, Hoeng J, et al.: Overview of BEL track: Extraction of complex relationships and their conversion to BEL. *BioCreative VI Workshop 2017. Proceedings*. 2017. 57-61 [Reference Source](#)

4. Yadav S, Ramesh S, Saha S, Ekbal A: Relation Extraction from Biomedical and Clinical Text: Unified Multitask Learning Framework. *IEEE/ACM Trans Comput Biol Bioinform.* 2020; **PP**. [PubMed Abstract](#) | [Publisher Full Text](#)

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and is the work technically sound?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Biomedical natural language processing

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 03 February 2021

<https://doi.org/10.5256/f1000research.27086.r77522>

© 2021 Verspoor K et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Yuan Li

School of Computing and Information Systems, The University of Melbourne, Melbourne, Australia

Karin M. Verspoor

¹ School of Computing and Information Systems, The University of Melbourne, Melbourne, Vic, Australia

² School of Computing Technologies, RMIT University, Melbourne, Australia

This paper compares the performance of a few pretrained models (BERT/BioBERT/RoBERTa x

base/large) with different classification heads (Linear, RBERT) and pre-processing strategies (PH/SSplit for linear and SpChar/MSpChar for RBERT) on 3 relationship extraction datasets (2 biomedical and 1 general).

The results show that BioBERT is slightly better than BERT on 2 biomedical datasets but comparable to RoBERTa large on all 3 datasets. These observations can be explained because BioBERT is more domain-specific than BERT but the training corpus of RoBERTa also includes bio-related text such as PubMed abstracts and scientific literature. It is also shown that using RBERT head leads to a substantial improvement on the general dataset but not on two biomedical datasets. In addition, it is observed that models trained on randomly permuted data can still achieve better than random performance on all 3 datasets, which indicates that tokens themselves alone are still informative for the relationship extraction task.

This paper is well-structured and easy to follow. Although no new model is proposed, extensive experiments have been done comparing existing models in different settings. We appreciate that the code is provided, and that statistical variation is measured in the presented results.

However, there are still a few concerns and places requiring clarification in the manuscript.

The reason for RoBERTa base's poor performance using RBERT head is still unclear. The hyper-parameter settings from the original RoBERTa paper are used, but as we know that some hyper-parameters, like the learning rate, are associated with the size of mini-batch and input, it is worth doing a hyper-parameter search for different tasks. This may lead the authors to either find the cause of RoBERTa's poor performance or exclude one possible source. Also, the authors could try to fine tune RoBERTa for a longer time, like 5 more epochs, to see if that will improve the performance.

Due to the poor performance of RoBERTa base, it is better to include BERT large in all experiments. Since BERT base works reasonably well, a comparison between BERT base and BERT large can provide more meaningful insights, particularly in the context of a claim that larger models may offset the value of domain-specific tailoring.

Although the authors state that their observation suggests that domain specific models may be of limited use given the tendency of recent model pretraining regimes to incorporate ever broader sets of data, it is hard to make blanket decisions based on this, like whether it is worthwhile to first pre-train a model on domain specific data for a specific task or directly use a pre-trained model. In practice, people may still need to try out both ways in specific contexts. Do the authors have more suggestions for practitioners?

The results of randomizing token order are not surprising because the relationship extraction task is defined as sentence classification and it is known that sentence classifiers that do not use any order information still work reasonably well in various NLP tasks ^{1,2}. This is because each word individually conveys a lot of information, unlike the case in computer vision where pixels are only meaningful when combined. Segment reordering has even been shown to be beneficial in certain tasks such as clinical Semantic Text Similarity ³, due to increasing robustness to linguistic variation. In addition, it is not a bad thing if a model can memorize training data that it has seen, and whether memorization will lead to poor generation is a separate question. So the authors should either provide evidence showing memorization will cause poor generalization in this case,

or try out some methods mentioned in the conclusion section to improve the performance.

Apart from the above points, this paper also lacks a related work section which is important to contextualize the contributions of the work, in particular since some of the conclusions contradict prior results. Some missing references are listed below, and particularly relevant is ⁴, which comes to different conclusions. Although some of these have been published close to or even after the submission date of this paper, it would still be worth adding some discussion of them to a revision.

Pre-trained models other than BERT/BioBERT/RoBERTa:

SpanBERT: Improving Pre-training by Representing and Predicting Spans (Transactions of ACL, 2020)

KnowBERT -- Knowledge Enhanced Contextual Word Representations (EMNLP 2019)

<https://www.aclweb.org/anthology/2020.emnlp-main.379>

<https://www.aclweb.org/anthology/2020.clinicalnlp-1.17>

Dedicated relation classification architecture:

Downstream Model Design of Pre-trained Language Model for Relation Extraction Task (arXiv 2020)

Relation classification as question answering:

Relation Extraction as Two-way Span-Prediction (arXiv 2020)

Error analysis of relation extraction dataset and models:

TACRED Revisited: A Thorough Evaluation of the TACRED Relation Extraction Task (ACL 2020)

Minor comments:

"knowledge entombed" -- We hardly think that the scientific literature is a tomb for knowledge. Perhaps find a different word.

"this represents practical issues" → "this introduces practical issues".

References

1. Arora: A Simple but Tough-to-Beat Baseline for Sentence Embeddings. *5th International Conference on Learning Representations, ICLR 2017*. 2017. [Reference Source](#)
2. Shen: Baseline Needs More Love: On Simple Word-Embedding-Based Models and Associated Pooling Mechanisms. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2018. [Reference Source](#)
3. Wang: Evaluating the Utility of Model Configurations and Data Augmentation on Clinical Semantic Textual Similarity. *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*. 2020. [Reference Source](#)
4. Gu: Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *arxiv:2007.15779*. 2020. [Reference Source](#)

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and is the work technically sound?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Partly

Competing Interests: KV receives funding from Elsevier BV for research on methods for information extraction from biochemical texts and has a related provisional patent.

Reviewer Expertise: Biomedical natural language processing

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research