OctaNLP: A Benchmark for Evaluating Multitask Generalization of Transformer-Based Pre-trained Language Models

Zakaria Kaddari, Youssef Mellah, Jamal Berrich, Mohammed G. Belkasmi, and Toumi Bouchentouf

Abstract In the last decade, deep learning based Natural Language Processing (NLP) models achieved remarkable performance on the majority of NLP tasks, especially, in machine translation, question answering and dialogue. NLP language models shifted from uncontextualized vector space models like word2vec and Glove in 2013, and 2014, to contextualized LSTM-based model like ELMO and ULMFit in 2018, to contextualized transformer-based models like BERT. Transformer-based language models are already trained to perform very well on individual NLP tasks. However, when applied to many tasks simultaneously, their performance drops considerably. In this paper, we overview NLP evaluation metrics, multitask benchmarks, and the recent transformer-based language models. We discuss the limitations of the current multitask benchmarks, and we propose our octaNLP benchmark for comparing the generalization capabilities of the transformer-based pre-trained language models on multiple downstream NLP tasks simultaneously.

Keywords NLP · Multitask · Benchmark · octaNLP · Metrics · Transformer

Z. Kaddari (✉) · Y. Mellah · J. Berrich · T. Bouchentouf
LaRSA Laboratory, AIRES Team, National School of Applied Sciences,
Université Mohammed Premier, Oujda, Morocco
e-mail: z.kaddari@ump.ac.ma

Y. Mellah
e-mail: y.mellah@ump.ac.ma

J. Berrich
e-mail: j.berrich@ump.ac.ma

T. Bouchentouf
e-mail: t.bouchentouf@ump.ac.ma

M. G. Belkasmi
SmartICT Laboratory, National School of Applied Sciences,
Université Mohammed Premier, Oujda, Morocco
e-mail: m.belkasmi@ump.ac.ma

S. Bennani et al. (eds.), *WITS 2020*, Lecture Notes in Electrical Engineering 745,
https://doi.org/10.1007/978-981-33-6893-4_21

1 Introduction

The rate of adoption of NLP applications by companies and customers is increasing rapidly. This is mostly due to the progress that has been made by deep learning (DL) and transformer-based pre-trained language models (LM) [21]. Some of these LM can even be used, and personalized directly without any knowledge of machine learning or coding.

The field of NLP contains many tasks, and new tasks are proposed each year by the NLP research community. In deep learning based NLP, some tasks are more studied than others. In the last couple of years, DL transformer-based LM achieved state-of-the-art performances on the majority of NLP tasks.

The field of NLP does not have a universal evaluation metric that can be used to evaluate the performance of new models on every task. But rather, a variety of metrics, like, BLEU [19], and ROUGE [15], among others. These metrics are used for specific tasks, BLEU is used in machine translation (MT) for example, and ROUGE for summarization. However, when we want to evaluate the generalization of a LM on multiple tasks at once, we confront with a major problem, which is the lack of a universal and unique metric for all or at least a subset of NLP tasks. This is one of NLP's open challenges [10] that is attracting more research in recent years. The study of this problem is the core of this paper, where we provide an overview of evaluation metrics and multitask NLP benchmarks along with our proposed octaNLP benchmarking approach for comparing the generalization capabilities of DL transformer-based language models.

This paper is organized as follows, the next section overviews the most used NLP evaluation metrics. Section 3 describes the available multitask NLP benchmarks. In Sect. 4, we overview the most important DL transformer-based pre-trained LM. In Sect. 5, we discuss the limitations of the available multitask NLP benchmarks, and we propose our octaNLP benchmark for comparing the generalization performance of transformer-based pre-trained LM on multiple downstream tasks simultaneously. Finally, we finish the paper with a conclusion.

2 Evaluation Metrics in NLP

In the field of NLP, there is no single metric that can be used to evaluate the performance of a system on all NLP tasks. But rather, a set of metrics that are used depending on the task. In the case of classification for example, the accuracy metric can be used, which indicates the percentage of correct classifications. Other metrics can also be used in the case of classification, like F1, exact match, and Matthews correlation coefficient [17] These classification metrics are not specific to NLP, but rather, used in a wide range of areas and disciplines. On the other hand, there are metrics that are specific to NLP, the most used ones are listed below:

Table 1 The most used metrics in NLP along with their associated tasks

Metric	Tasks
Accuracy	Question answering
	Sentiment analyses
	Paraphrasing
	Natural language inference
	Coreference resolution
	Word sense disambiguation
F1	Paraphrasing
	Natural language inference
	Question answering
Exact match	Question answering
Matthews correlation coefficient	Grammatical acceptability
BLEU	Machine translation
ROUGE	Summarization
Perplexity	Language modeling
Pearson correlation coefficient	Sentence similarity
Spearman correlation coefficient	Sentence similarity

BLEU: The bilingual evaluation understudy (BLEU) [19] is an automatic metric that was initially defined to evaluate systems for machine translation (MT). However, it is now also used in other Natural Language Generation (NLG) tasks like, summarization, and dialogue. The BLEU score is used to compare a candidate translation to one or more reference translations. This score can range between O and 1, for 1 being a perfect translation. BLEU has many strong advantages; it is an automatic metric, language independent, and proved to correlate highly with human judgment.

ROUGE: Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [15] is a set of metrics that are used to evaluate the performance of automatic summarization or machine translation systems. ROUGE metrics compare a candidate summary or translation to one or more reference summarizations or translations.

In Table 1, we list the most used metrics in NLP along with their associated tasks.

3 Multitask NLP Benchmarks

decaNLP: The Natural Language Decathlon (decaNLP) [18] benchmark was introduced in 2018. The goal of this benchmark is to evaluate single models that can generalize to many different NLP tasks simultaneously. The tasks included in the benchmark are, semantic parsing, natural language inference, question answering, document summarization, machine translation, sentiment analysis, semantic role labeling, goal-oriented dialogue, pronoun resolution, and relation extraction. All

Table 2 GLUE tasks along with their associated datasets and metrics

Task	Dataset	Metric
Question answering	QNLI [23]	Accuracy
Sentiment analyses	SST-2 [25]	Accuracy
Paraphrasing	MRPC [7] and QQP[a]	F1/accuracy
Grammatical acceptability	CoLA [30]	Matthews correlation coefficient [17]
Sentence similarity	STS-B [1]	Pearson/Spearman correlation coefficients
Natural language inference	MNLI [31] and RTE [5]	F1/accuracy
Coreference resolution	WNLI [13]	Accuracy

[a] https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs

these tasks were framed as a question answering problem, and are trained jointly. All training instances are in the form of (question, context, answer) triplets. To be able to evaluate the generalization of NLP models across all tasks simultaneously, the creators of decaNLP defined their own score that they called decaScore, which is simply the sum of the scores of all tasks. The creators of decaNLP also provided and evaluated three baseline models, a pointer-generator sequence-to-sequence (S2S) model [24], an S2S model augmented with self-attentive encoder and decoder layers [26], and an S2S model augmented with a coattention mechanism [32]. In addition to the three baseline models, the creators of decaNLP also built their own model that they called the multitask question answering network (MQAN). MQAN learns all of decaNLP tasks jointly, and does not require any task-specific modules or parameters. This model achieved improved performance on the majority of decaNLP tasks.

GLUE: Similar to decaNLP, the General Language Understanding Evaluation (GLUE) [28][1] benchmark aims to drive research in general NLP models that can generalize well to a variety of different tasks. However, the scope of GLUE is more limited than decaNLP, because GLUE is only concerned with Natural Language Understanding (NLU) tasks. These tasks along with their associated datasets and metrics are listed in Table 2. To evaluate the general performance of NLP models across all tasks, GLUE define a single score, which is simply the average score on all tasks with all tasks having the same weight. For tasks with multiple metrics, the benchmarking algorithm of GLUE first averages those metrics to get a single task score. Since its release, a large number of models have been tested on the benchmark, especially, transformer-based pre-trained language models. Recent models have surpassed the human performance on GLUE for the majority of its tasks.

SuperGLUE: Like GLUE, The SuperGLUE [27] [2] benchmark aims to evaluate general NLP models on a variety of tasks simultaneously. This benchmark was introduced after the surpassing of human performance on GLUE by the recent models on

[1] https://gluebenchmark.com.

[2] https://super.gluebenchmark.com.

Table 3 Super GLUE tasks along with their associated datasets and metrics

Task	Dataset	Metric
Question answering	BoolQ [2], COPA [8], MultiRC [11], ReCoRD [33]	Accuracy/F1/EM
Natural language inference	CB [16] and RTE [5]	Accuracy/F1
Word sense disambiguation	WiC [20]	Accuracy
Coreference resolution	WSC[a]	Accuracy

[a] https://cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WS.html

the majority of GLUE tasks, which made GLUE no longer suitable for tracking the progress towards general NLU models. SuperGLUE differs from GLUE in that it contains more difficult and challenging NLU tasks with more diverse tasks formats. SuperGLUE adopts the same scoring philosophy as GLUE, by weighting each task equally and averaging all tasks score's, to provide a single general score. The tasks used in SuperGLUE along with their associated datasets and metrics are listed in Table 3.

SentEval: SentEval [4] is a benchmark and a toolkit for evaluating the quality of universal general-purpose sentence representations. The goal of this benchmark is to drive research in finding sentence representations that can yield good results when applied on a variety of different downstream NLP tasks. SentEval contains a diverse set of tasks including, binary and multi-class classification, entailment and semantic relatedness, Semantic Textual Similarity (STS), paraphrase detection, caption-Image retrieval, and sentiment analyses.

4 Transformer-Based Pre-trained Language Models

In this section, we will provide a short overview of the most important transformer-based pre-trained language models. Most of these models are based on BERT, the first transformer-based pre-trained language model released. Figure 1 shows the models that were derived from BERT, along with what was added to them.

BERT [6]: a pre-trained model based on the transformer model. BERT is designed to perform deep two-way representations from unlabeled text by jointly conditioning the left and right context in all the layers. It pre-trains a next sentence prediction task to understand sentence relationships.

SemBERT [34]: This model is capable of explicitly absorbing contextual semantics over a BERT backbone. SemBERT keeps the convenient usability of its BERT precursor with light fine-tuning and without substantial task-specific modifications. Compared with BERT, SemBERT is as simple in concept but more powerful. It obtains new state-of-the-art or substantially improves results on ten reading comprehension and language inference tasks.

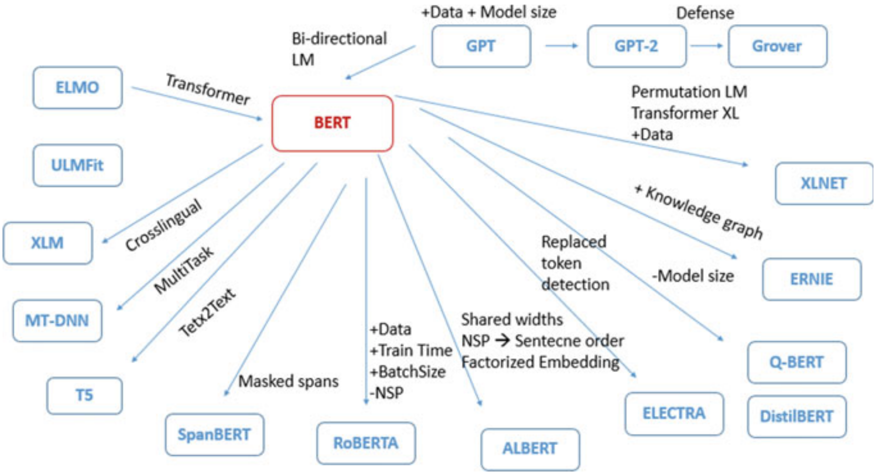

Fig. 1 Models that were derived from BERT, along with what was added to them

StructBERT [29]: This model was made by incorporating language structures into pre-training. Specifically, its trained with two auxiliary tasks to make the most of the sequential order of words and sentences, which leverage language structures at the word and sentence levels, respectively.

ALBERT [12]: This model presents two parameter-reduction techniques to lower memory consumption and increase the training speed of BERT: Splitting the embedding matrix into two smaller matrices and using repeating layers split among groups.

ELECTRA [3]: is a new pre-training approach which trains two transformer models: the generator and the discriminator. The generator's role is to replace tokens in a sequence, and is therefore trained as a masked language model. The discriminator, which is the model we're interested in, tries to identify which tokens were replaced by the generator in the sequence.

T5 [22]: is an encoder-decoder model pre-trained on a multi-task mixture of unsupervised and supervised tasks and for which each task is converted into a text-to-text format.

BART [14]: This model combines bidirectional and auto-regressive transformers. It is a denoising autoencoder built with a sequence-to-sequence model that can tackle a wide range of NLP tasks from NLU to NLG. Although it is particularly effective when fine-tuned for text generation tasks. BART achieved new state-of-the-art on numerous tasks such as dialogue, question answering, and summarization.

5 Our OctaNLP Benchmarking Approach

In Sect. 3, we reviewed the available multitask NLP benchmarks. We saw that decaNLP include 10 diverse tasks, with different evaluation metrics, such as F1, accuracy, BLEU and ROUGE. The variety of tasks and metrics makes decaNLP a perfect benchmark for evaluating the generalization of NLP models. However, decaNLP was released before BERT, the first transformer-based pre-trained language model. Therefore, it is not known if the benchmark is compatible with those kind of models. To the date of the writing of this paper, and to the best of our knowledge, no transformer-based pre-trained language model has been tested on decaNLP.

As for GLUE and SuperGLUE, we saw in the same section, that these two models are only concerned with evaluating the generalization on NLU tasks. The lack of any NLG task such as machine translation, summarization or dialogue, inhibits these two benchmarks from evaluating the generalization capabilities on all NLP tasks.

As for SentEval benchmark, its only goal is to evaluate the generalization of sentence representations, and same as GLUE and SuperGLUE, it is only concerned with NLU tasks.

To overcome the limitations of these multitask benchmarks, we propose a novel benchmark that we call octaNLP for evaluating the generalization capabilities of transformer-based pre-trained language models. The 8 tasks that we included in our benchmark covers the two pillars of NLP, NLU and NLG. Therefore, we think that our benchmark is more suitable for evaluating the generalization capabilities across all NLP tasks.

In Table 4, we list the datasets that we considered in our benchmark, along with their associated tasks and evaluation metrics.

We defined a single overall score, that we called octaScore, which is simply the average of the scores of all 8 tasks. We applied our benchmarking approach to two of the recent transformer-based pre-trained language models, BART and T5. Table 5 shows the results of these two models on each individual task along with the overall octaScore. We plan to apply our octaNLP benchmark to other models as a future work.

Table 4 The datasets adopted in octaNLP, along with their associated tasks and evaluation metrics

Dataset	Task	Metric
SQuAD [23]	Question answering	F1
MNLI [31]	Natural language inference	F1
QQP[a]	Semantic textual similarity	Accuracy
QNLI [23]	Question answering	Accuracy
RTE [5]	Natural language inference	Accuracy
MRPC [7]	Paraphrasing	Accuracy
CoLA [30]	Grammatical acceptability	Matthews correlation coefficient (Mcc)
CNN/DM [9]	Summarization	Rouge

[a] https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs

Table 5 Benchmarking results of BART and T5 on octaNLP benchmark

Model	SQuAD	MNLI	QQP	QNLI	RTE	MRPC	CoLA	CNN/DM	Octa score
	F1	m	Acc	Acc	Acc	Acc	Mcc	R1	
BART	94.6	89.9	**92.5**	94.9	87.0	90.4	62.8	**44.16**	82.03
T5	**95.64**	**92.0**	90.4	**96.7**	**92.5**	89.2	**70.8**	43.52	**83.84**

From this experiment, we can see that the T5 model achieved the best octaScore, which means that this model can generalize better than BART on diverse NLP tasks. This is because T5 is a text-to-text model, meaning that it approaches every NLP task in the same manner, as text input to text output.

6 Conclusion and Future Work

Transformer-based pre-trained language models have achieved remarkable results on many individual NLP tasks, but are still lacking generalization capabilities to be applied to multiple tasks simultaneously. In this paper, we provided an overview of NLP evaluation metrics, multitask benchmarks, and transformer-based pre-trained language models. We presented the limitations of the current multitask benchmarks, and proposed our octaNLP benchmark for comparing the generalization capabilities of the transformer-based pre-trained language models on multiple downstream NLP tasks simultaneously. As a future work, we plan to test the multitask generalization capabilities of other transformer-based pre-trained language models using our octaNLP benchmark.

References

1. Cer D, Diab M, Agirre E, Lopez-Gazpio I, Specia L (2017) SemEval-2017 task 1: semantic textual similarity multilingual and crosslingual focused evaluation. In: Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017). Association for Computational Linguistics, Vancouver, Canada, pp 1–14
2. Clark C, Lee K, Chang MW, Kwiatkowski T, Collins M, Toutanova K (2019) BoolQ: exploring the surprising difficulty of natural yes/no questions. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, vol 1. Association for Computational Linguistics, Minneapolis, Minnesota, pp 2924–2936
3. Clark K, Luong MT, Le QV, Manning CD (2020) ELECTRA: pre-training text encoders as discriminators rather than generators. arXiv e-prints arXiv:2003.10555
4. Conneau A, Kiela D (2018) SentEval: an evaluation toolkit for universal sentence representations. In: Proceedings of the eleventh international conference on language resources and evaluation (LREC (2018) European Language Resources Association (ELRA). Miyazaki, Japan

5. Dagan I, Glickman OMB (2006) Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment. Lecture notes in computer science, Vol 3944. Springer
6. Devlin J, Chang M, Lee K, Toutanova K (2018) BERT: pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805
7. Dolan WB, Brockett C (2005) Automatically constructing a corpus of sentential paraphrases. In: Proceedings of the third international workshop on paraphrasing (IWP2005)
8. Gordon A, Kozareva Z, Roemmele M (2012) SemEval-2012 task 7: choice of plausible alternatives: An evaluation of commonsense causal reasoning. In: *SEM 2012: the first joint conference on lexical and computational semantics, volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the sixth international workshop on semantic evaluation (SemEval 2012). Association for Computational Linguistics, Montréal, Canada, pp 394–398
9. Hermann KM, Kociský T, Grefenstette E, Espeholt L, Kay W, Suleyman M, Blunsom P (2015) Teaching machines to read and comprehend. CoRR abs/1506.03340
10. Kaddari Z, Mellah Y, Berrich J, Belkasmi MG, Bouchentouf T (2021) Natural language processing: Challenges and future directions. In: Masrour T, El Hassani I, Cherrafi A (eds) Artificial intelligence and industrial applications. Springer International Publishing, Cham, pp 236–246
11. Khashabi D, Chaturvedi S, Roth M, Upadhyay S, Roth D (2018) Looking beyond the surface: a challenge set for reading comprehension over multiple sentences. In: Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies (Long Papers), vol 1. Association for Computational Linguistics, New Orleans, Louisiana, pp 252–262
12. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R (2019) ALBERT: a lite bert for self-supervised learning of language representations. arXiv e-prints arXiv:1909.11942
13. Levesque HJ, Davis E, Morgenstern L (2011) The winograd schema challenge. In: AAAI spring symposium: logical formalizations of commonsense reasoning, volume 46 (2011)
14. Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, Stoyanov V, Zettlemoyer L (2019) BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv e-prints arXiv:1910.13461
15. Lin CY (2004) ROUGE: A package for automatic evaluation of summaries. Text summarization branches out. Association for Computational Linguistics, Barcelona, Spain, pp 74–81
16. Marneffe M-C, Simmons M, Tonhauser J (2019) The commitmentbank: investigating projection in naturally occurring discourse. https://ojs.ub.uni-konstanz.de/sub/index.php/sub/article/view/601
17. Matthews B Comparison of the predicted and observed secondary structure of t4 phage lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Struct 405(2) 442 – 451
18. McCann B, Keskar NS, Xiong C, Socher R (2018) The natural language decathlon: multitask learning as question answering. CoRR abs/1806.08730
19. Papineni K, Roukos S, Ward T, Zhu WJ (2002) Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the association for computational linguistics. Association for Computational Linguistics, Philadelphia, Pennsylvania, pp 311–318
20. Pilehvar MT, Camacho-Collados J (2019) WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, vol 1. Association for computational linguistics, Minneapolis, Minnesota, pp 1267–1273
21. Qiu X, Sun T, Xu Y, Shao Y, Dai N, Huang X (2020) Pre-trained models for natural language processing: a survey. arXiv e-prints arXiv:2003.08271
22. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ (2019) Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv e-prints arXiv:1910.10683
23. Rajpurkar P, Zhang J, Lopyrev K, Liang P (2016) SQuAD: 100,000+ questions for machine comprehension of text. In: Proceedings of the 2016 conference on empirical methods in natural language processing. Association for computational linguistics, Austin, Texas, pp 2383–2392

24. See, A., Liu, P.J., Manning, C.D.: Get to the point: Summarization with pointer-generator networks. CoRR abs/1704.04368 (2017)
25. Socher R, Perelygin A, Wu J, Chuang J, Manning CD, Ng A, Potts C (2013) Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the 2013 conference on empirical methods in natural language processing. Association for Computational Linguistics, Seattle, Washington, pp 1631–1642
26. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser LU, Polosukhin I (2017) Attention is all you need. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in neural information processing systems, vol 30. Curran Associates, Inc., pp 5998–6008
27. Wang A, Pruksachatkun Y, Nangia N, Singh A, Michael J, Hill F, Levy O, Bowman SR (2019) Superglue: a stickier benchmark for general-purpose language understanding systems
28. Wang A, Singh A, Michael J, Hill F, Levy O, Bowman SR (2019) Glue: a multi-task benchmark and analysis platform for natural language understanding
29. Wang W, Bi B, Yan M, Wu C, Bao Z, Xia J, Peng L, Si L (2019) StructBERT: incorporating language structures into pre-training for deep language understanding. arXiv e-prints arXiv:1908.04577
30. Warstadt A, Singh A, Bowman SR (2018) Neural network acceptability judgments. CoRR abs/1805.12471
31. Williams A, Nangia N, Bowman S (2018) A broad-coverage challenge corpus for sentence understanding through inference. In: Proceedings of the 2018 conference of the North American Chapter of the association for computational linguistics: human language technologies, vol 1 (Long Papers). Association for Computational Linguistics, New Orleans, Louisiana, pp 1112–1122
32. Xiong C, Zhong V, Socher R (2016) Dynamic coattention networks for question answering. CoRR abs/1611.01604
33. Zhang S, Liu X, Liu J, Gao J, Duh K, Durme BV (2018) Record: bridging the gap between human and machine commonsense reading comprehension. CoRR abs/1810.12885
34. Zhang Z, Wu Y, Zhao H, Li Z, Zhang S, Zhou X, Zhou X (2019) Semantics-aware BERT for language understanding. arXiv e-prints arXiv:1909.02209