Oğuzhan Karahan, et. al. International Journal of Engineering Research and Applications
www.ijera.com
ISSN: 2248-9622, Vol. 10, Issue 9, (Series-VI) September 2020, pp. 50-54

RESEARCH ARTICLE OPEN ACCESS

The Evaluation of Natural Language Understanding Models for Answering Dialogue Questions

Oğuzhan Karahan*, Ahmet Gürhanlı**
**(Program of Computer Engineering, Institute of Graduate Study, Istanbul Aydin University, Istanbul, Turkey*
*** (Department of Computer Engineering, Faculty of Engineering, Istanbul Aydin University, Istanbul, Turkey*

ABSTRACT
With the development of technology and the creation of Transformer-based models, beneficial improvements have been achieved in the field of natural language understanding (NLU). The most important and primary task for the development of natural language understanding is reading comprehension which contains various tasks that are accepted from datasets for named entity recognition and intent classification. At the same time, we offer datasets for the banking, investment and social domains that are encountered very often in daily life. To guarantee shared evaluation scheme, we created models that simplify different NLU tasks, and contain datasets from diverse domains and applications. In conclusion, we offer a general estimation with some standard baselines and future in Transformer-based models of Turkish language.
Keywords: Turkish, natural language understanding, answering dialogues, deep learning, evaluation

Date of Submission: 25-09-2020 Date of Acceptance: 07-10-2020

I. INTRODUCTION

The field of natural language understanding has undergone a considerable progress information reusability. The same improvements took domain in transfer learning. This situation has been a known to all phenomenon in the domain of computer vision. This progress was achieved with the recent advent of healthy, general purpose language models proper for configurable. Like BERT (Devlin et al., 2019), DistilBERT (Sanh et al., 2020), ALBERT (Lan et al., 2020) and DIET (Bunk et al., 2020).

These released models have provided many improvements in the NLU field and accelerated the latest developments. In light of these developments, progress has accelerated significantly, and different models of Transformer-based method have been released frequently to further develop the latest technology.

For progress, a widely available evaluation environment is first required. And then this needs to be standardized. These will establish the NLU standards. The furthest popular NLU standard is GLUE (Wang et al., 2019a), which was introduced very recently. GLUE basically has skills such as answering questions, finding relationships between different texts, and analyzing emotions. It has been observed that some tasks need a lot of training examples while others have finite training data. In addition, the training set and the test set show different areas for some tasks. This encourages models learn universal language representations and perform transferring knowledge across several tasks and domains. GLUE is built on current datasets, and GLUE's main supports are careful task selection. In addition to providing an evaluation platform, it displays the results in a dashboard.

NLU has made great progress for some languages. The first of these is English and the second is Chinese. Because these languages are richer in terms of both pre-trained models and evaluation criteria compared to other languages. In this article, since there is no Turkish language support in GLUE tasks, data creation and understanding measurement are provided through the data sets we have created and data sets for investment, banking and social areas have been created. GLUE tasks have been carefully selected to cover different features of the language. Following the GLUE model, we adjusted the tasks to fit text classification to reduce the model evaluation stages.

Our main contributions in this study are listed below:
1. We have developed a platform to evaluate and present model results on the dashboard.
2. INVD: We offer a new intent classification for the investment domain.
3. BNKD: We offer a new intent classification task for the banking domain.
4. SCLD: We offer a new intent classification task for the social domain.
5. Evaluation of baseline is done using Transformer-based models for Turkish language.

The content of the study is as follows; we described the study in Chapter 2. We explained the tasks in Chapter 3. In Chapter 4 we refer to the basic methods. In Chapter 5, we apply and evaluate all datasets and models. Lastly, we offer the conclusion in Chapter 6.

II. RELATED WORK

An important part of progress; it's the evaluation of natural language understanding models. There are many factors to consider for new models. However, there is no clear standard on how to choose the model. Testing models will show the right choice. This situation has revealed multitasking comparisons between models.

An example of this standard is SentEval (Conneau and Kiela, 2018). This is a model for evaluating the quality of sentence placement consisting of seventeen tasks. In addition, 10 tasks are provided to identify which language features are used with sentence placement. In each task, language models take single or double sentence insertion as input and solve a classification or regression problem.

Another example of this standard is decaNLP (McCann et al., 2018). This model has 10 tasks and the selection of tasks is much more varied compared to SentEval. All tasks convert into an automatic question-answering format, which is the core task of the NLU.

Lastly, The General Language Understanding Evaluation (GLUE) is the important example of benchmark. (Wang et al., 2019a). This model suggests 9 available, well-structured tasks. More varied and difficult tasks should be chosen so that the comparison does not resemble SentEval.

All models and criteria given above are limited to English. Efforts to provide support for many languages include the MNLI (Williams et al., 2018) dataset with support for 14 languages and the XNLI dataset (Conneau et al., 2018). Similarly, the model of the SQuAD dataset (Rajpurkar et al., 2016) with 10 language support is XQuAD (Artetxe et al., 2019).

Unfortunately, the examples mentioned are limited to some languages. There are no studies for the Turkish language. There is no standard for comparing language models in Turkish. In this paper, we try the intent classification in three different domain datasets. These datasets include banking, investment and social tasks.

Table 1: Presented datasets and evaluation metrics.

Name	Train	Domain	Metrics
BNKD	6.5k	Banking	F1-Score
INVD	3.9k	Investment	F1-Score
SCLD	2.4k	Social	F1-Score

III. TASKS

In this study, we focused on tasks with not very large data sets. The small relative smallness of the dataset may require external knowledge to solve the model's problems. Our main goal is to create a general model for Turkish language support. We explain the details of the tasks in the sections below. We provide summary information in Table 1.

3.1 BNKD

We introduce a new intent classification dataset, named BNKD, extracting 6.5k dataset from banking domain. There are at least 6 training data for each intent. It contains 419 intents in total. The task is to predict the classification of a given banking domain data sets.

3.2 INVD

We introduce a new intent classification dataset, named INVD, extracting 3.9k dataset from investment domain. There are at least 3 training data for each intent. It contains 134 intents in total. The task is to predict the classification of a given investment domain data sets.

3.3 SCLD

We introduce a new intent classification dataset, named SCLD, extracting 2.4k dataset from social domain. There are at least 3 training data for each intent. It contains 122 intents in total. The task is to predict the classification of a given social domain data sets.

IV. BASELINE

Recent studies show that Bidirectional Encoder Representations (BERT) models created on the base of Transformer architecture work well. These are the results evaluated considering the GLUE benchmarks. These models are pre-trained using the Masked Language Model (MLM). In this section we will explain these language models: BERT (Devlin et al., 2019), Distilbert (Sanh et al., 2020), Albert (Lan et al., 2020) and DIET (Bunk et al., 2020). The specified Transformer-based models are used in the most current form at the time of writing this paper.

In order to evaluate models correctly, we need to provide similar parameters for each task.

Transformers (Wolf et al., 2019) library was used for model training. Input parameters for all models are listed in Table 3.

4.1 BERT

While creating the language model, Transformer architecture was taken as basis. It performs training based on Next Sentence Prediction (NSP) as well as the Masked Language Model (MLM).

It is a Turkish language model that includes the Turkish corpus of Wikipedia. The model used is Turkish Cased BERT. This pre-trained model consists of 32k words. WordPiece (Wu et al., 2016) is used as a tokenizer.

4.2 DISTILBERT

As the name suggests, it has the same architecture as BERT, but the number of layers is 2 times less. The processes used in transformer architecture are generally optimized and developed within the framework of modern linear algebra.

The Turkish Cased DistilBERT model used is a Turkish-based model containing all Wikipedia Turkish corpora. This pre-trained model consists of 32k words. WordPiece (Wu et al., 2016) is used as a tokenizer.

4.3 ALBERT

The architecture of this language model is similar to that of BERT in that it uses a non-linear GELU and a Transformer encoder. In addition, ALBERT also uses the two parameters reduction technique that removes some of the barriers in pre-trained models.

The Turkish uncased ALBERT model used is a Turkish-based model containing all Wikipedia Turkish corpora. This pre-trained model consists of 18k words. WordPiece (Wu et al., 2016) is used as a tokenizer.

4.4 DIET

DIET (Dual Intent and Entity Transformer) language model is multitasking architecture for intent classification and entity recognition. One of the most important features of this language model is its ability to combine words with pre-trained language model.

Adding pre-trained words to language models improves the overall accuracy of all tasks. This language model is the highest performing model and is six times faster to train.

V. EVALUATION

Knowing that there is no exact measure of comparing models, models should be compared according to their average performance. Since the difficulty levels of the models are different, it would be fair to compare all models to their average performance.

Table 2: Baseline evaluation on Transformer-based models. AVG is the weighted average of the tasks evaluated.

Model	AVG	BNKD	INVD	SCLD
BERT	81.4	74.9	90.1	79.4
DistilBERT	82.6	76.4	**91.2**	80.3
ALBERT	82.7	77.6	90.8	79.7
DIET	**85.0**	**81.4**	89.5	**84.2**

In comparison to on the pre-trained BERT models, DIET model performs slightly better. The exception is the small training datasets. When the results are examined, the lack of pre-trained BERT models is shown. In the comparison made over the data sets created, the DIET model looks better on average.

Based on pre-trained models, the corpus plays a very important role in the performance of the task. Pre-trained models for different areas may not always be the right choice. This is because the question answering task is Wikipedia based and the BERT model is Wikipedia corpus trained.

When evaluating the results, it is necessary to know that the DIET model is not a pre-trained model. However, it did obtain better on average than other Transformer architecture models. The DIET model performed best in 2 out of 3 tasks. Moreover, DIET has the smallest performance gap between BNKD and SCLD.

This shows that it can work more generally in different domains. When the results of other language models are examined, the average score is very similar. The DIET model performed the worst in the INVD dataset, while DistilBERT performed the best in the same dataset.

DIET language model should also be tested on different domains in Turkish. And it should be compared with other language models. In this paper, it may have performed best in 2 out of 3 tasks. However, different language models can perform better in different domains. With a single language model, it may not be possible to show the best performance on every domain. The way to get good results with few data sets for different domains may be to use or create domain-based embeddings.

VI. CONCLUSION

In this study, we reviewed Turkish-sourced Transformer-based models, which do not have widely documented documents. The aim is to

develop and popularize Turkish NLU models. Task selection is very important for success in models. We focused on different text lengths and types to make the model more successful. In conclusion, different models have proven to perform better for different tasks, as there is no single model that works best for all tasks.

In our work, we evaluated the DIET model with other pre-trained transformer-based language models for the Turkish language. We found that the DIET model was successful in 2 out of 3 tasks and also performed the best on average.

We aimed to provide a common assessment for different tasks of different models. For that purpose, many existing resources had to be adapted and datasets has been created in different domains.

In such model comparisons, model performance should be focused at, knowing that each model will give different results for each problem. Data such as the status of the pre-trained model, size of the trained model, training speed and input parameters can be taken as a performance measure.

With the development of pre-trained language models, it will become increasingly difficult to compare and choose language models. For Turkish language, it is necessary to establish a standard for the comparison of language models by creating data sets in different domains. As future work, we aim to create new data sets and standardize them.

REFERENCES

[1]. M. Artetxe, S. Ruder, and D. Yogatama. 2019. *On the cross-lingual transferability of monolingual representations.*

[2]. P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606.*

[3]. A. Conneau and D. Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018).*

[4]. A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

[5]. J. Devlin, M. Chang, K. Lee, and K. Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

[6]. E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018).*

[7]. J. Howard and S. Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339.

[8]. D.P. Kingma and J. Ba. 2014. Adam: A method for stochastic optimization. Cite *arxiv: 1412.6980* Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

[9]. G. Lample and A. Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

[10]. P. Lison and J. Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

[11]. B. McCann, N.S. Keskar, C. Xiong, and R. Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730.*

[12]. B.H.R. Sennrich and A. Birch. 2016. Neural machine translation of rare words with subword units. *In Association for Computational Linguistics (ACL), pages 1715–1725*.

[13]. A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S.R. Bowman. 2019a. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In the Proceedings of ICLR.

[14]. A. Wang, I.F. Tenney, Y. Pruksachatkun, K. Yu, J. Hula, P. Xia, R. Pappagari, S. Jin, R.T. McCoy, R. Patel, Y. Huang, J. Phang, E. Grave, H. Liu, N. Kim, P.M. Htut, T. F'evry,

B. Chen, N. Nangia, A. Mohananey, K. Kann, S. Bordia, N. Patry, D. Benton, E. Pavlick, and S.R. Bowman. 2019b. jiant 1.2: A software toolkit for research on general purpose text understanding models.

[15]. A. Williams, N. Nangia, and S. Bowman. 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, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

[16]. P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

[17]. P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

[18]. A. Conneau, R. Rinott, G. Lample, A. Williams, S.R. Bowman, H. Schwenk, and V. Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natu- ral Language Processing*. Association for Computational Linguistics.

[19]. T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

[20]. A. Williams, N. Nangia, and S. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In NAACL.

[21]. Y. Wu, M. Schuster, Z. Chen, Q.V Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between hu- man and machine translation. *arXiv preprint arXiv:1609.08144*.

[22]. T. Bunk, D. Varshneya, V. Vlasov, and A. Nichol. 2020. DIET: Lightweight language understanding for dialogue systems. *CoRR, abs/2004.09936*.

[23]. Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942, 2019*.

[24]. V. Sanh, L. Debut, J. Chaumond, and T. Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108, 2019*.

Table 3: Input parameters for all models are given as follows.

Parameter	Value
N-Grams	[3,5]
Classifier Epochs	50
Transformer Layers	4
Transformer Size	256
Masked Language Model	True
Drop Rate	0.25
Weight Sparsity	0.7
Batch Size	[64, 256]
Embedding Dimension	30
Hidden Layer Sizes (for text)	[512, 128]