Proceedings of the 16th Conference on Computer Science and Intelligence Systems pp. 115–119
DOI: 10.15439/2021F7
ISSN 2300-5963 ACSIS, Vol. 25

Evaluation of Neural Network Transformer Models for Named-Entity Recognition on Low-Resourced Languages

Ridewaan Hanslo
University of Pretoria
Gauteng, South Africa
Email: ridewaan.hanslo@up.ac.za

***Abstract*—Neural Network (NN) models produce state-of-the-art results for natural language processing tasks. Further, NN models are used for sequence tagging tasks on low-resourced languages with good results. However, the findings are not consistent for all low-resourced languages, and many of these languages have not been sufficiently evaluated. Therefore, in this paper, transformer NN models are used to evaluate named-entity recognition for ten low-resourced South African languages. Further, these transformer models are compared to other NN models and a Conditional Random Fields (CRF) Machine Learning (ML) model. The findings show that the transformer models have the highest F-scores with more than a 5% performance difference from the other models. However, the CRF ML model has the highest average F-score. The transformer model's greater parallelization allows low-resourced languages to be trained and tested with less effort and resource costs. This makes transformer models viable for low-resourced languages. Future research could improve upon these findings by implementing a linear-complexity recurrent transformer variant.**

I. Introduction

XLM-Roberta (XLM-R) is a recent transformer model that has reported state-of-the-art results for Natural Language Processing (NLP) tasks and applications, such as Named-Entity Recognition (NER), Part-of-Speech (POS) tagging, phrase chunking, and Machine Translation (MT) [2], [8]. The NER and POS sequence tagging tasks have been extensively researched [1]-[6], [8], [9]. However, within the past few years, the introduction of new Deep Learning (DL) transformer model architectures such as XLM-R, Multilingual Bidirectional Encoder Representations from Transformers (M-BERT) and Cross-Lingual Language Model (XLM) lowers the time needed to train large datasets through greater parallelization [7]. This allows low-resourced languages to be trained and tested with less effort and resource costs, with state-of-the-art results for sequence tagging tasks [1], [2], [8]. M-BERT as a single language model pre-trained from monolingual corpora performs very well with cross-lingual generalization [10]. Furthermore, M-BERT is capable of capturing multilingual representations [10]. On the other hand, XLM pre-training has led to strong improvements on NLP benchmarks [11]. Additionally, XLM models have contributed to significant improvements in NLP studies involving low-resource languages [11]. These transformer models are usually trained on very large corpora with datasets in terabyte (TB) sizes.

A recent study by [1] researched the "*Viability of Neural Networks for Core Technologies for Resource-Scarce Languages*". These resource-scarce languages are ten of the 11 official South African (SA) languages, with English being excluded. The languages are considered low-resourced, with Afrikaans (af) being the more resourced of the ten [1], [9]. This recent study looked at sequence tagging (POS tagging and NER) and sequence translation (Lemmatization and Compound Analysis), comparing two Bidirectional Long Short-Term Memory with Auxiliary Loss (bi-LSTM-aux) NN models to a baseline Conditional Random Fields (CRF) model. The annotated data used for the experiments are derived from the National Centre for Human Language Technology (NCHLT) text project. The results suggest that NN architectures such as bi-LSTM-aux are viable for NER and POS tagging tasks for most SA languages [1]. However, within the study by [1], NN did not outperform the CRF Machine Learning (ML) model. Rather they advised further studies be conducted using NN transformer models on resource-scarce SA languages. For this reason, this study builds upon the previous study, using the XLM-R DL architecture. Therefore, the purpose of this study is to evaluate the performance of the NLP NER sequential task using two XLM-R transformer models. In addition, the experiment results are compared to previous research findings.

A. Research Questions

RQ_1 – How does the XLM-R neural network transformer models perform with NER on the low-resourced SA languages using annotated data?

RQ_2 – How does the XLM-R transformer models compare to other neural network and machine learning models with NER on the low-resourced SA languages using annotated data?

B. Paper Layout

The remainder of this paper comprises of the following sections: Sect. II provides information on the languages and

datasets; Sect. III presents the language model architecture. The experiment settings are presented in Sect. IV and the results and a discussion of the research findings are provided in Sect. V. Section VI concludes the paper with the limitations of this study and recommendations for further research.

II. Languages and Datasets

As mentioned by [1], SA is a country with at least 35 spoken languages. Of those languages, 11 are granted official status. The 11 languages can further be broken up into three distinct groups. The two West-Germanic languages, English and Afrikaans (af). Five disjunctive languages, Tshivenda (ve), Xitsonga (ts), Sesotho (st), Sepedi (nso) and Setswana (tn) and four conjunctive languages, isiZulu (zu), isiXhosa (xh), isiNdebele (nr) and Siswati (ss). A key difference between SA disjunctive and conjunctive languages is the former has more words per sentence than the latter. Therefore, disjunctive languages have a higher token count than conjunctive languages. For further details on conjunctive and disjunctive languages with examples, see [1].

The datasets for the ten evaluated languages are available from the South African Centre for Digital Language Resources online repository (https://repo.sadilar.org/). These annotated datasets are part of the NCHLT Text Resource Development Project, developed by the Centre for Text Technology (CTexT, North-West University, South Africa) with contributions by the SA Department of Arts and Culture. The annotated data is tokenized into five phrase types. These five phrase types are:

1. ORG - Organization
2. LOC - Location
3. PER - Person
4. MISC - Miscellaneous
5. OUT - not considered part of any named-entity

The datasets consist of SA government domain corpora. Therefore, the SA government domain corpora are used to do the experiments and comparisons. Eiselen [9] provides further details on the annotated corpora.

III. Language Model Architecture

XLM-Roberta (XLM-R) is a transformer-based multilingual masked language model [2]. This language model trained on 100 languages uses 2.5 TB of CommonCrawl (CC) data [2]. From the 100 languages used by the XLM-R multilingual masked language model, it is noted that Afrikaans (af) and isiXhosa (xh) are included in the pre-training.

The benefit of this model, as indicated by [2] is, training the XLM-R model on cleaned CC data increases the amount of data for low-resource languages. Further, because the XLM-R multilingual model is pre-trained on many languages, low-resource languages improve in performance due to positive transfer [2].

Conneau et al. [2] reports the state-of-the-art XLM-R model performs better than other NN models such as M-BERT and XLM on question-answering, classification, and sequence labelling.

Two transformer models are used for NER evaluation. The XLM-R_{Base} NN model and the XLM-R_{Large} NN model. The XLM-R_{Base} model has 12 layers, 768 hidden states, 12 attention heads, 250 thousand vocabulary size, and 270 million parameters. The XLM-R_{Large} model has 24 layers, 1024 hidden states, 16 attention heads, 250 thousand vocabulary size, and 550 million parameters [2]. Both pre-trained models are publicly available (https://bit.ly/xlm-rbase, https://bit.ly/xlm-rlarge).

IV. Experimental Settings

The experimental settings for the XLM-R_{Base} and XLM-R_{Large} models are described next, followed by the evaluation metrics and the corpora descriptive statistics.

A. XLM-R Settings

The training, validation, and test dataset split was 80%, 10%, and 10%, respectively. Both pre-trained models were fine-tuned with the following experimental settings:

1. Training epochs: 10
2. Maximum sequence length: 128
3. Learning rate: 0.00006
4. Training batch size: 32
5. Gradient accumulation steps: 4
6. Dropout: 0.2

B. Evaluation Metrics

Precision, Recall and F-score are evaluation metrics used for text classification tasks, such as NER. These metrics are used to measure the model's performance during the experiments. The formulas for these metrics leave out the correct classification of true negatives (*tn*) and false negatives (*fn*), referred to as negative examples, with greater importance placed on the correct classification of positive examples such as true positives (*tp*) and false positives (*fp*) [12]. For example, correctly classified spam emails (*tp*) are more important than correctly classified non-spam emails (*tn*). In addition, multi-class classification was used for the research experiments to classify a token into a discrete class from three or more classes. The metric's macro-averages were used for evaluation and comparison. Macro-averaging (*M*) treats classes equally, while micro-averaging (μ) favors bigger classes [12]. Each evaluation metric and its formula as described by [12] are listed below. (*M*) treats classes equally, while micro-averaging (μ) favors bigger classes [12]. Each evaluation metric and its formula as described by [12] are listed below.

Precision$_M$: "*the number of correctly classified positive examples divided by the number of examples labeled by the system as positive"* (1).

$$\frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i + fp_i}}{l} \tag{1}$$

Recall$_M$: "*the number of correctly classified positive examples divided by the number of positive examples in the data"* (2).

$$\frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i + fn_i}}{l} \quad (2)$$

Fscore$_M$: "*a combination of the above"* (3).

$$\frac{(\beta^2 + 1)Precision_M Recall_M}{\beta^2\, Precision_M + Recall_M} \quad (3)$$

C. Corpora Descriptive Statistics

Table I provides descriptive statistics for the language's training data.

V. Results and Discussion

A. Results

Table II displays the precision scores of the two XLM-R transformer models compared to models used by [1] and [9]. The Afrikaans (af) language has the highest precision score in this comparison, with 81.74% for the XLM-R_{Large} model. The XLM-R_{Base} model has the lowest overall score of 38.59% for the Sesotho (st) language. The CRF model has the highest precision scores for six of the ten languages, including the highest average score of 75.64%. The bold scores in Table II, III and IV show the highest evaluation metric score for each language and the model with the highest average score.

Table III displays the recall scores for the ten low-resourced SA languages. As with the precision evaluation metric, the Afrikaans (af) language has the highest recall score, with 87.07% for the XLM-R_{Large} model. The XLM-R_{Base} model has the lowest recall score of 39.41% for the Sesotho (st) language. The CRF and bi-LSTM-aux models have the highest recall scores for three of the ten languages, respectively, with the latter model having the highest average score of 72.48%.

Table IV displays the F-score comparison. The Afrikaans (af) language produced the highest F-score, with an 84.25% for the XLM-R_{Large} model. The XLM-R_{Base} model has the lowest F-score of 38.94% for the Sesotho (st) language. The CRF model has the highest F-score for four of the ten languages, including the highest average score of 73.22%.

B. Discussion

The two research questions are answered in this section. The first question is on the transformer model's performance using the three-evaluation metrics, whereas the second question compares the transformer model's performance to the CRF and bi-LSTM models used in the previous SA NER studies.

TABLE I.
THE TEN LANGUAGES TRAINING DATA DESCRIPTIVE STATISTICS

Language	Writing System	Tokens	Phrase Types
Afrikaans (af)	Mixed	184 005	22 693
isiNdebele (nr)	Conjunctive	129 577	38 852
isiXhosa (xh)	Conjunctive	96 877	33 951
isiZulu (zu)	Conjunctive	161 497	50 114
Sepedi (nso)	Disjunctive	161 161	17 646
Sesotho (st)	Disjunctive	215 655	18 411
Setswana (tn)	Disjunctive	185 433	17 670
Siswati (ss)	Conjunctive	140 783	42 111
Tshivenda (ve)	Disjunctive	188 399	15 947
Xitsonga (ts)	Disjunctive	214 835	17 904

RQ_1 – How does the XLM-R neural network transformer models perform with NER on the low-resourced SA languages using annotated data?

The XLM-R_{Large} and XLM-R_{Base} transformer models produced F-scores that ranged from 39% for the Sesotho (st) language to 84% for the Afrikaans (af) language. Further, many of the models recall scores were greater than 70% whereas the precision scores were averaging at 65%. Remember, in this instance, the recall metric emphasizes the average per-named-entity effectiveness of the classifier to identify named-entities, whereas, the precision metric compares the alignment of the classifier's average per-named-entities to the named-entities in the data. All F-scores were above 60% except the Sesotho language, which for both XLM-R models were below 40%. The reason for the low F-scores of the Sesotho (st) language has not been identified, however, it is posited that an investigation into using different hyper-parameter tuning and dataset splits can produce higher F-scores. Sesotho (st) is clearly the outlier during the experiments. For instance, the Sesotho (st) language exclusion from the transformer models results moves the average F-score from 67% to 71%. For the low-resourced SA languages, this is a notable improvement.

RQ_2 – How does the XLM-R transformer models compare to other neural network and machine learning models with NER on the low-resourced SA languages using annotated data?

The transformer models were also compared to the findings of previous studies. In particular, [9] used a CRF ML model to do NER sequence tagging on the ten resource-scarce SA languages. Further, [1] implemented bi-LSTM-aux NN models, both with and without embeddings on the same dataset. When analyzing the F-scores, the CRF model has the highest F-scores for four of the ten languages, and the bi-LSTM-aux models shared four of the highest F-scores equally (see Table IV). Meanwhile, the XML-R transformer models have two of the highest F-scores (see Table IV). Although, the transformer models were the only models to produce F-scores greater than 80% for the Afrikaans (af)

TABLE II.
THE PRECISION % COMPARISON BETWEEN TRANSFORMER MODELS AND PREVIOUS SA LANGUAGE NER STUDIES

Precision					
	CRF*	**bi-LSTM-aux****	**bi-LSTM-aux emb****	**XLM-R_{Base}**	**XLM-R_{Large}**
af	78.59%	73.61%	73.41%	79.15%	**81.74%**
nr	77.03%	**78.58%**	n/a***	74.06%	73.43%
xh	**78.60%**	69.83%	69.08%	64.94%	65.97%
zu	**73.56%**	72.43%	73.44%	71.10%	71.91%
nso	76.12%	75.91%	72.14%	**77.23%**	n/a****
st	**76.17%**	53.29%	50.31%	38.59%	39.34%
tn	**80.86%**	74.14%	73.45%	67.09%	68.73%
ss	69.03%	**70.02%**	69.93%	65.39%	65.99%
ve	**73.96%**	67.97%	63.82%	58.85%	60.61%
ts	**72.48%**	72.33%	71.03%	63.58%	63.58%
Average	**75.64%**	70.81%	68.51%	65.99%	65.70%

* As reported by [9]. ** As reported by [1]. *** No embeddings were available for isiNdebele.
**** The model was unable to produce scores for Sepedi.

TABLE III.
THE RECALL % COMPARISON BETWEEN TRANSFORMER MODELS AND PREVIOUS SA LANGUAGE NER STUDIES

Recall					
	CRF*	**bi-LSTM-aux****	**bi-LSTM-aux emb****	**XLM-R_{Base}**	**XLM-R_{Large}**
af	73.32%	78.23%	78.23%	86.16%	**87.07%**
nr	73.26%	**79.20%**	n/a***	78.51%	78.02%
xh	**75.61%**	73.30%	72.78%	63.53%	64.74%
zu	66.64%	72.64%	74.32%	74.23%	**74.58%**
nso	72.88%	79.66%	77.63%	**80.59%**	n/a****
st	**70.27%**	55.56%	57.73%	39.41%	39.71%
tn	75.47%	**77.42%**	74.71%	73.39%	76.22%
ss	60.17%	71.44%	**72.82%**	70.09%	70.97%
ve	**72.92%**	65.91%	67.09%	63.24%	64.22%
ts	69.46%	**71.44%**	71.25%	68.34%	69.40%
Average	71.00%	**72.48%**	71.84%	69.74%	69.43%

* As reported by [9]. ** As reported by [1]. *** No embeddings were available for isiNdebele.
**** The model was unable to produce scores for Sepedi.

TABLE IV.
THE F-SCORE % COMPARISON BETWEEN TRANSFORMER MODELS AND PREVIOUS SA LANGUAGE NER STUDIES

F-score					
	CRF*	**bi-LSTM-aux****	**bi-LSTM-aux emb****	**XLM-R_{Base}**	**XLM-R_{Large}**
af	75.86%	75.85%	75.74%	82.47%	**84.25%**
nr	75.10%	**78.89%**	n/a***	76.17%	75.60%
xh	**77.08%**	71.52%	70.88%	63.58%	64.68%
zu	69.93%	72.54%	**73.87%**	72.54%	73.17%
nso	74.46%	77.74%	74.79%	**78.86%**	n/a****
st	**73.09%**	54.40%	53.77%	38.94%	39.48%
tn	**78.06%**	75.74%	74.07%	69.78%	71.91%
ss	64.29%	70.72%	**71.35%**	67.57%	68.34%
ve	**73.43%**	66.92%	65.41%	60.68%	61.99%
ts	70.93%	**71.88%**	71.14%	65.57%	66.12%
Average	**73.22%**	71.62%	70.11%	67.61%	67.28%

* As reported by [9]. ** As reported by [1]. *** No embeddings were available for isiNdebele.
**** The model was unable to produce scores for Sepedi.

language. This is a significant improvement for NER research on the SA languages.

The comparative analysis identified the Sesotho (st) language as the lowest-performing language across the studies, albeit the CRF model has an F-score of 73%, making it an outlier. If the Sesotho (st) language is excluded from the evaluation, then the metric scores for the transformer models begin to look much different.

For example, the highest average recall score of 72.48% by [1] belonged to the bi-LSTM-aux model, yet, the XLM-R_{Large} model, with Sesotho excluded, was able to produce an average recall score of 73.15%. Similarly, with Sesotho excluded, the average F-score and precision score were 71% and 69%, respectively, which are close to the high scores of the previous studies.

This study reveals that the NN transformer models perform fairly well on low-resource SA languages with NER sequence tagging, and Afrikaans (af) outperforms the other languages using these models. During the NN transformer model experiments, the disjunctive languages had a higher token count, while conjunctive languages had a higher phrase type count (see Table I). However, there is no distinct performance difference between individual disjunctive and conjunctive languages both during the XLM-R experiments and when compared to the other NN and ML models. Nonetheless, except for the CRF model, conjunctive languages had a higher F-score average than disjunctive languages, even with the disjunctive Sesotho (st) language excluded.

The Sesotho (st) language is a clear outlier in this study, with the CRF baseline model F-score being 33% more than the XLM-R models and 18% more than the bi-LSTM-aux models. Interestingly, while both the isiXhosa (xh) and Afrikaans (af) languages were included in the pre-training of the XLM-R model (see Section III) isiXhosa (xh) underperformed when compared to the CRF and bi-LSTM-aux models. This finding suggests including a language in the XLM-R model pre-training does not guarantee good performance during evaluation. It is posited that the experiment results could be improved upon. For instance, additional fine-tuning of the hyper-parameters for each NN model can be done per language, given the available resources. Further, in agreement with [9], the annotation quality could be a contributor to the performance of the models.

VI. Limitations and Further Research

The limitations of this research are the lack of resource capacity to apply additional hyperparameter optimizations on the transformer models per language. Additionally, the named entities of the corpora would need to be investigated and re-evaluated. It is posited, that the quality of the annotations could be improved upon, and the dataset could be re-evaluated using an updated list of named entities.

Additional research, therefore, could implement the transformer models with discrete fine-tuning parameters per language to produce higher F-scores. In addition, the transformer models could be used to evaluate other NLP sequence tagging and sequence-to-sequence tasks such as POS tagging, Phrase chunking, and MT on the low-resource SA languages. Finally, sequence tagging tasks could be evaluated using a linear-complexity recurrent transformer variant.

References

[1] M. Loubser, and M. J. Puttkammer, "Viability of neural networks for core technologies for resource-scarce languages". *Information*, Switzerland, 2020. https://doi.org/10.3390/info11010041

[2] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, *Unsupervised Cross-lingual Representation Learning at Scale*, 2020. https://doi.org/10.18653/v1/2020.acl-main.747

[3] B. Plank, A. Søgaard, and Y. Goldberg, "Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss". *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Short Papers*, 2016. https://doi.org/10.18653/v1/p16-2067

[4] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition". *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference*, 2016. https://doi.org/10.18653/v1/n16-1030

[5] J. Lafferty, A. McCallum, and C. N. F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data". *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*, 2001 https://doi.org/10.29122/mipi.v11i1.2792

[6] E. D. Liddy, "Natural Language Processing. In Encyclopedia of Library and Information Science". In *Encyclopedia of Library and Information Science*, 2001.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need". *Advances in Neural Information Processing Systems*, 2017.

[8] M. A. Hedderich, D. Adelani, D. Zhu, J. Alabi, U. Markus, and D. Klakow, *Transfer Learning and Distant Supervision for Multilingual Transformer Models: A Study on African Languages*, 2020. https://doi.org/10.18653/v1/2020.emnlp-main.204

[9] R. Eiselen, "Government domain named entity recognition for South African languages". *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, 2016.

[10] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual BERT?" *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 2020. https://doi.org/10.18653/v1/p19-1493

[11] A. Conneau, and G. Lample, "Cross-lingual language model pretraining". *Advances in Neural Information Processing Systems*, 2019.

[12] M. Sokolova, and G. Lapalme, "A systematic analysis of performance measures for classification tasks". *Information Processing and Management*, *45*(4), 2009. https://doi.org/10.1016/j.ipm.2009.03.002