MphayaNER: Named Entity Recognition for Tshivenḓa

Rendani Mbuvha[*,1,2], David I. Adelani[3], Tendani Mutavhatsindi[2], Tshimangadzo Rakhuhu[4], Aluwani Mauda[5], Tshifhiwa Joshua Maumela[6], Andisani Masindi, Seani Rananga[7], Vukosi Marivate[7,8], and Tshilidzi Marwala[9]

[1] School of Electronic Engineering and Computer Science, Queen Mary University of London.
[2] School of Statistics and Actuarial Science, University of Witwatersrand
[3] Department of Computer Science, University College London.
[4] School of Business, National College of Ireland
[5] Centre for Development Support, University of the Free State
[6] School of Electrical and Electronic Engineering, University of Johannesburg
[7] Department of Computer Science, University of Pretoria
[8] Lelapa AI
[9] United Nations University
[*] Corresponding email `r.mbuvha@qmul.ac.uk`

Abstract

Named Entity Recognition (NER) plays a vital role in various Natural Language Processing tasks such as information retrieval, text classification, and question answering. However, NER can be challenging, especially in low-resource languages with limited annotated datasets and tools. This paper adds to the effort of addressing these challenges by introducing MphayaNER, the first Tshivenḓa NER corpus in the news domain. We establish NER baselines by *fine-tuning* state-of-the-art models on MphayaNER. The study also explores zero-shot transfer between Tshivenḓa and other related Bantu languages, with Setswana, chiShona and Kiswahili showing the best results. Augmenting MphayaNER with Setwana data was also found to improve model performance significantly. Both MphayaNER and the baseline models are made publicly available [1].

Named Entity Recognition (NER) is a natural language processing (NLP) task that involves identifying and classifying named entities such as person names, location names, and organization names in text (Mohit, 2014; Eiselen, 2016). NER is critical in various NLP applications, including information retrieval, text classification, and question answering. Identifying named entities is challenging, particularly in low-resource languages where annotated datasets and tools for NER are limited (Adelani et al., 2021). Tshivenḓa is a minority language spoken by approximately 1.3 million[2] people, predominantly in South Africa and its northern neighbours of Zimbabwe, Mozambique and Botswana StatsSA (2016). Given the relatively small number of native speakers of the language, Tshivenḓa has received limited attention in the NLP community, especially concerning NER. NER activity in Tshivenḓa has been limited to SADiLaR corpus, which is primarily restricted to named entities in the government domain (Eiselen, 2016). The task of NER in Tshivenḓa is complicated by the fact that Tshivenḓa, like many other Bantu languages, is a morphologically rich language with complex morphological forms that can pose challenges for NER systems (van Huyssteen & Griesel, 2014; Adelani et al., 2022). Additionally, Tshivenḓa has limited resources, including annotated datasets and pre-trained models, which are essential for developing NER systems. These challenges highlight the need to develop NER systems tailored to Tshivenḓa, specifically designed to handle the language's morphological richness and resource-poor nature. We address the underrepresentation of Tshivenḓa in NER by *fine-tuning* state-of-the-art pre-trained language models and further investigate cross-lingual zero-shot transfer between Tshivenḓa and five other African languages.

[1] `https://github.com/rendanim/MphayaNER`
[2] `http://salanguages.com/Tshivenda/index.htm`

TSHIVENḒA

Tshivenḓa is part of the South-Eastern group of Bantu languages and is known for its similarities to the Karanga Dialect of Shona, as noted in (Cassimjee, 1992). Tshivenḓa employs the Latin alphabet along with five letters that have accents. Four of these letters are dental consonants with a circumflex accent beneath them (ḓ, ḽ, ṋ, ṱ). One is a dot above the letter for the velar (ṅ). Seven vowels are represented using five vowel letters. J and Q are used exclusively for foreign words and names. Tshivenḓa has a rich morphological structure featuring 21 noun classes (Musehane, 2014). Like many Bantu languages, it only uses two tones - high (denoted by an acute accent) and low (no diacritic) (Reynolds, 1997). The population of Tshivenḓa speakers is the second least in South Africa after isiNdebele speakers. As part of the minority speakers, there has been little attention to ensuring this language is included in digital platforms, particularly news media.

MPHAYANER CORPUS

We use data from Vuk'uzenzele [3], a monthly newspaper published in multiple languages, including Tshivenḓa, by the communications unit of the South African Government. For this study, Tshivenḓa Vuk'uzenzele news articles were used. The articles were originally in PDF format, they were converted to text files using py2PDF (Fenniak et al., 2022), cleaned and standardised by the Data Science for social impact (DSFSI) research group at the University of Pretoria Marivate et al. (2023); Lastrucci et al. (2023)[4]. We manually preprocess the converted text file to minimise text extraction errors such as repeated phrases and erroneous word concatenation or splitting to create the MphayaNER corpus.

ANNOTATION

We make use of the same annotation guideline, scheme and tags as done by (Adelani et al., 2021). We also make use of ELISA annotation tool for labelling Tshivenḓa texts. We recruited 5 native speakers for the labelling and they were trained using the MasakhaNER guideline for labeling personal name (PER), organization (ORG), location (LOC), and dates (DATE). To speed up annotation effort, we assigned a sentence to be annotated by only two annotators. After annotation, we make use of the adjudication function of the ELISA tool to detect mistakes between annotators. Annotators met to resolve conflict, and for sentences that still had disagreement, we ask the lead Tshivenḓa annotator to decide the most appropriate tag for the token. Our new dataset is known as MphayaNER. After annotation and verification, we divided the 1,359 sentences into TRAIN, DEV and TEST split based on 70%/10%/20% ratio. In total, the dataset consists of 40,778 tokens, where TRAIN, DEV and TEST splits have about 29,239, 4,212 and 7,327 tokens respectively as shown in Table 1. In MphayaNER, there are few entities in general, we found only 251 PER, 409 LOC, 702 ORG , and 504 DATE entities.

BASELINE MODELS AND RESULTS

We provide baseline NER models based on *fine-tuning* multilingual pre-trained language models (PLMs) – this is the state-of-the-art technique. We compared fine-tuning mDeBERTaV3 (He et al., 2021), one of the current state-of-the-art PLMs and two Africa-centric PLMs, i.e. AfriBERTa(Ogueji et al., 2021) and AfroXLMR Alabi et al. (2022). We benchmark on several sizes of the AfroXLMR model (small, base, and large versions). All models are trained for 20 epochs, a learning rate of $5e-5$.

Table 1(a) shows the result on MphayaNER; the AfriBERTa model has the worst result (58.7 F1) since Tshivenḓa and other related Southern Bantu languages were not covered during pre-training. mDeBERTaV3 had an impressive result (66.4 F1) despite seeing only two Bantu languages during pre-training(Kiswahili and isiXhosa). AfroXLMR has seen more Bantu languages, about eight. We found that the more the capacity/size of AfroXLMR increases, the better the performance — AfroXLMR-small (66.1 F1), AfroXLMR-small (67.3 F1), and AfroXLMR-large (71.0 F1). We found all the models struggle to identify DATE in the corpus.

[3] https://www.vukuzenzele.gov.za/

[4] https://github.com/dsfsi/vukuzenzele-nlp/

Split	# of sentences	# of tokens
TRAIN	951	29,239
DEV	135	4,212
TEST	273	7,327

(a) Data split for MphayaNER

Model	F1-score DATE	LOC	ORG	PER	AVG
mDeBERTaV3	39.2	76.6	67.6	96.2	66.4
AfriBERTa	37.4	60.2	54.8	84.2	58.7
AfroXLMR-small	41.0	77.4	62.4	**99.4**	66.1
AfroXLMR-base	43.4	**80.2**	62.6	97.0	67.3
AfroXLMR-large	**47.6**	77.2	**71.0**	98.8	**71.0**

(b) Baseline results for MphayaNER. Average over 5 runs.

Table 1: Data split and Benchmark results (F1-score) for MphayaNER.

lang.	F1-score LOC	ORG	PER	AVG
src lang. → tgt lang.				
sna	47.6	30.6	**46.0**	40.0
swa	52.4	**37.2**	40.6	43.2
tsn	**66.4**	31.4	45.0	**46.2**
xho	57.0	29.3	35.3	38.4
zul	50.8	29.8	32.6	37.0
same lang.: *govt → news*				
ven	53.4	33.6	20.8	35.2

(a) Zero-shot eval. on MphayaNER

src lang.	F1-score DATE	LOC	ORG	PER	AVG
zero-shot eval.					
sna	64.6	49.2	30.2	38.8	42.9
swa	**72.6**	32.6	37.4	38.8	43.2
tsn	56.0	52.4	29.6	39.4	43.7
co-training eval.					
sna+ven	50.0	82.4	72.8	97.0	73.5
swa+ven	48.0	81.2	75.4	**97.4**	73.5
tsn+ven	50.0	**86.2**	**76.8**	94.6	74.5
sna+swa+tsn+ven	52.6	86.0	75.6	93.0	**74.9**

(b) zero-shot and co-training results on all named entity tags

Table 2: Transfer learning and co-training evaluation results on MphayaNER.

Transfer result We examine how other Bantu languages adapt to Tshivenḓa in the zero-shot setting. Similarly, we compared the adaptation of a NER model trained on several Southern African languages in MasakhaNER 2.0 corpus (Adelani et al., 2022) tTshivenḓa. The source transfer languages are chiShona (sna), Kiswahili(swa), Setswana(tsn), isiXhosa (xho), and isiZulu (zul). Also, we perform a zero-shot adaptation of SADILAR Tshivenḓa NER corpus based on government data to MphayaNER corpus based on the news corpus (*govt → news*). We excluded the miscellaneous tag (MISC) and DATE tag for the cross-domain experiment since the MphayaNER dataset does not have MISC.

Table 2 shows the result of the cross-lingual transfer with or without the DATE tag. Our result in Table 2(a) shows a significant drop in performance when transferring to Tshivenḓa for all languages. Surprisingly, the government domain transfer (35.2 F1) to MphayaNER is even worst than language adaptation from other Southern African languages ($> 37.0F1$), which shows that the difference in domain has a big effect on transfer performance. Furthermore, we performed transfer experiments, including the DATE tags for the top three languages with the best transfer results in Table 2(a). We found out that Kiswahili and chiShona have very good transfer results for the DATE tag (over 64.0 F1); the performance for Kiswahili was particularly unexpected since it is not a Southern Bantu language like others. Setswana (tsn) seems to be better on LOC entity since native speakers of the languages are geographically close (i.e both are spoken in the same region in the northeastern parts of South Africa). All the source languages struggle to adapt to both ORG and PER entities. Lastly, we performed co-training experiments by combining the MphayaNER corpus with the one of chiShona, Kiswahili and Setswana; we found that co-training improves the performance by $2.5-3.9$ F1 points. In general, only co-training with Setswana leads to 3.5 F1 points.

CONCLUSION

In this paper, we present first Tshivenḓa NER corpus in the news domain based on the Vuk'zenzele dataset. We establish baselines for this task by fine-tuning multilingual pre-trained state-of-the-art language models. We also investigate zero-shot transfer between Tshivenḓa and related Bantu languages finding that Setswana and Kiswahili give the best results. Additional experiments show augmenting MphayaNER with chiShona data significantly improves model performance.

REFERENCES

David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Mboning Tchiaze Elvis, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo Lerato Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Oluwaseun Adeyemi, Gilles Quentin Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu, and Dietrich Klakow. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 4488–4508, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL `https://aclanthology.org/2022.emnlp-main.298`.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, et al. Masakhaner: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131, 2021.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 4336–4349, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL `https://aclanthology.org/2022.coling-1.382`.

Farida Cassimjee. *An autosegmental analysis of Venda tonology*. Taylor & Francis, 1992.

Roald Eiselen. Government domain named entity recognition for south african languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 3344–3348, 2016.

Mathieu Fenniak, Matthew Stamy, pubpub zz, Martin Thoma, Matthew Peveler, exiledkingcc, and PyPDF2 Contributors. The PyPDF2 library, 2022. URL `https://pypi.org/project/PyPDF2/`.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2021.

Richard Lastrucci, Isheanesu Dzingirai, Jenalea Rajab, Andani Madodonga, Matimba Shingange, Daniel Njini, and Vukosi Marivate. Preparing the vuk'uzenzele and za-gov-multilingual south african multilingual corpora. *arXiv preprint arXiv: Arxiv-2303.03750*, 2023.

Vukosi Marivate, Daniel Njini, Andani Madodonga, Richard Lastrucci, and Isheanesu Dzingirai. The vuk'uzenzele south african multilingual corpus, February 2023. URL `https://doi.org/10.5281/zenodo.7598539`.

Behrang Mohit. *Named Entity Recognition*, pp. 221–245. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014. ISBN 978-3-642-45358-8. doi: 10.1007/978-3-642-45358-8_7. URL `https://doi.org/10.1007/978-3-642-45358-8_7`.

Prof Nelson Musehane. An analysis of common methods of number distinction in tshivenda. *IOSR Journal of Humanities and Social Science*, 19(3):06–12, 2014. doi: 10.9790/0837-19360612.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pp. 116–126, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.mrl-1.11. URL `https://aclanthology.org/2021.mrl-1.11`.

William T Reynolds. Post-high tone shift in venda nominals. *Ms., University of the Witwatersrand*, 1997.

StatsSA. Community survey 2016, 2016. URL https://www.statssa.gov.za/?page_id=3839.

Gerhard B van Huyssteen and Marissa Griesel. Translation technology in south africa. In *Routledge Encyclopedia of Translation Technology*, pp. 365–374. Routledge, 2014.