Transfer Learning and Distant Supervision for Multilingual Transformer Models: A Study on African Languages

Michael A. Hedderich[1], David I. Adelani[1], Dawei Zhu[1], Jesujoba Alabi[1,2], Udia Markus[3] & Dietrich Klakow[1]
[1]Saarland University, Saarland Informatics Campus, Germany
[2]DFKI GmbH, Saarbrücken, Germany [3]Nuhu Bamalli Polytechnic, Zaira, Nigeria
{mhedderich,didelani,dzhu,dietrich.klakow}@lsv.uni-saarland.de
jesujoba_oluwadara.alabi@dfki.de

Abstract

Multilingual transformer models like mBERT and XLM-RoBERTa have obtained great improvements for many NLP tasks on a variety of languages. However, recent works also showed that results from high-resource languages could not be easily transferred to realistic, low-resource scenarios. In this work, we study trends in performance for different amounts of available resources for the three African languages Hausa, isiXhosa and Yorùbá on both NER and topic classification. We show that in combination with transfer learning or distant supervision, these models can achieve with as little as 10 or 100 labeled sentences the same performance as baselines with much more supervised training data. However, we also find settings where this does not hold. Our discussions and additional experiments on assumptions such as time and hardware restrictions highlight challenges and opportunities in low-resource learning.

1 Introduction

Deep learning techniques, including contextualized word embeddings based on transformers and pretrained on language modelling, have resulted in considerable improvements for many NLP tasks. However, they often require large amounts of labeled training data, and there is also growing evidence that transferring approaches from high to low-resource settings is not straightforward. In (Loubser and Puttkammer, 2020a), rule-based or linguistically motivated CRFs still outperform RNN-based methods on several tasks for South African languages. For pretraining approaches where labeled data exists in a high-resource language, and the information is transferred to a low-resource language, Hu et al. (2020) find a significant gap between performance on English and the cross-lingually transferred models. In a recent study, Lauscher et al. (2020) find that the transfer for multilingual transformer models is less effective for resource-lean settings and distant languages. A popular technique to obtain labeled data quickly and cheaply is distant and weak supervision. Kann et al. (2020) recently inspected POS classifiers trained on weak supervision. They found that in contrast to scenarios with simulated low-resource settings of high-resource languages, in truly low-resource settings this is still a difficult problem. These findings also highlight the importance of aiming for realistic experiments when studying low-resource scenarios.

In this work, we analyse multilingual transformer models, namely mBERT (Devlin et al., 2019; Devlin, 2019) and XLM-RoBERTa (Conneau et al., 2019). We evaluate both sequence and token classification tasks in the form of news title topic classification and named entity recognition (NER). A variety of approaches have been proposed to improve performance in low-resource settings. In this work, we study (i) transfer learning from a high-resource language and (ii) distant supervision. We selected these as they are two of the most popular techniques in the recent literature and are rather independent of a specific model architecture. Both need auxiliary data. For transfer learning, this is labeled data in a high-resource language, and for distant supervision, this is expert insight and a mechanism to (semi-)automatically generate labels. We see them, therefore, as orthogonal and depending on the scenario and the data availability, either one or the other approach might be applicable.

Our study is performed on three, linguistically different African languages: Hausa, isiXhosa and Yorùbá. These represent languages with millions of users and active use of digital infrastructure, but with only very limited support for NLP technologies. For this aim, we also collected three new

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 2580–2591,
November 16–20, 2020. ©2020 Association for Computational Linguistics

datasets that are made publicly available alongside the code and additional material.[1]

We show both challenges and opportunities when working with multilingual transformer models evaluating trends for different levels of resource scarcity. The paper is structured into the following questions we are interested in:

- How do more complex transformer models compare to established RNNs?
- How can transfer-learning be used effectively?
- Is distant supervision helpful?
- What assumptions do we have to consider when targeting a realistic treatment of low-resource scenarios?

2 Languages and Datasets

In this work, we evaluate on three African languages, namely Hausa, isiXhosa and Yorùbá. Hausa is from the Afro-Asiatic family while isiXhosa and Yorùbá belong to different branches of the large Niger-Congo family. Hausa and Yorùbá are the second and third most spoken languages in Africa, and isiXhosa is recognized as one of the official languages in South Africa and Zimbabwe. Yorùbá has been part of the unlabeled training data for the mBERT multilingual, contextual word embeddings. Texts in Hausa and isiXhosa have been part of the XLM-RoBERTa training.

The three languages have few or no labeled datasets online for popular NLP tasks like named entity recognition (NER) and topic classification. We use the NER dataset by Eiselen (2016) for isiXhosa and the one by Alabi et al. (2020) for Yorùbá. We collected and manually annotated a NER dataset for Hausa and news title topic classification datasets for Hausa and Yorùbá. Table 1 gives a summary of the datasets. More information about the languages, the datasets and their creation process can be found in the Appendix.

Dataset Name	**Data Source**	**Full Train/ Val/ Test sentences**
Hausa NER*	VOA Hausa	1,014 / 145 / 291
Hausa Topic Class.*	VOA Hausa	2,045 / 290 /582
isiXhosa NER (Eiselen, 2016)	SADiLaR	5,138 / 608 / 537
Yorùbá NER (Alabi et al., 2020)	GlobalVoices	816 / 116 / 236
Yorùbá Topic Class.*	BBC Yoruba	1,340 / 189 / 379

Table 1: Datasets Summary. *Created for this work.

3 Experimental Settings

To evaluate different amounts of resource-availability, we use subsets of the training data with increasing sizes from ten to the maximally available number of sentences. All the models are trained on their corresponding language-model pre-training. Except if specified otherwise, the models are not fine-tuned on any other task-specific, labeled data from other languages. We report mean F1-score on the test sets over ten repetitions with standard error on the error bars. Additional experimental details are given in the following sections and the Appendix. The code is made publicly available online as well as a table with the scores for all the runs.

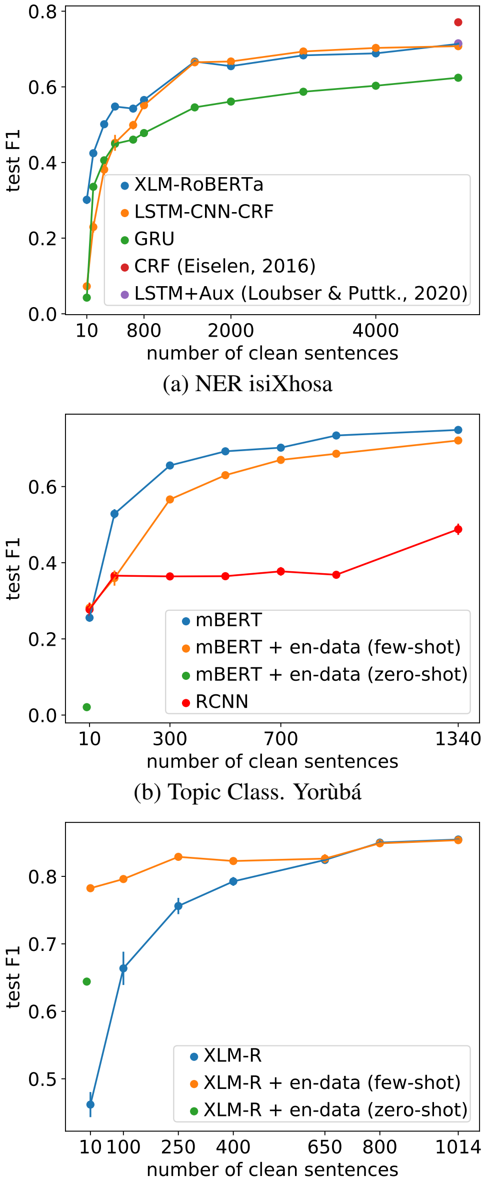
4 Comparing to RNNs

Loubser and Puttkammer (2020a) showed that models with comparatively few parameters, like CRFs, can still outperform more complex, neural RNNs models for several task and low-resource language combinations. This motivates the question whether model complexity is an issue for these low-resource NLP models. We compare to simple GRU based (Cho et al., 2014) models as well as the popular (non-transformer) combination of LSTM-CNN-CRF (Ma and Hovy, 2016) for NER and to the RCNN architecture (Lai et al., 2015) for topic classification. For these models, we use pre-trained, non-contextual word embeddings trained for the specific language. Figures 1a+b show that an increase in model complexity is not an issue in these experiments. For Hausa and Yorùbá and for the low resource settings for isiXhosa, BERT and XLM-RoBERTa actually outperform the other baselines, possibly due to the larger amounts of background knowledge through the language model pre-training. For larger amounts of task-specific training data, the LSTM-CNN-CRF and the transformer models perform similarly. One should note that for isiXhosa, the linguistically motivated CRF (Eiselen, 2016) still outperforms all approaches on the full dataset.

5 Transfer Learning

The mBERT and XLM-RoBERTa models are trained with tasks that can be obtained from unlabeled text, like masked language modelling. Addi-

[1] https://github.com/uds-lsv/transfer-distant-transformer-african



(a) NER isiXhosa

(b) Topic Class. Yorùbá

(c) Transfer Learn NER Hausa

Figure 1: Comparing to RNNs (a+b) and using transfer learning (b+c). Additional plots in the Appendix.

tionally, the multilingual models can be fine-tuned on task-specific, supervised data but from a different, high-resource language. There is evidence that the multilingual transformer models can learn parallel concepts across languages (Pires et al., 2019; Wu and Dredze, 2019; Hu et al., 2020). This allows to then apply or evaluate the model directly without having been fine-tuned on any labeled data in the target language (zero-shot) or on only a small amount of labeled data in the target language (few-shot).

For NER, we pre-train on the English CoNLL03 NER dataset (Tjong Kim Sang and De Meulder, 2003). For topic classification, the models are pre-trained on the English AG News corpus (Zhang et al., 2015). The texts in the high-resource English and the low-resource Hausa and Yorùbá target datasets share the same domain (news texts). One issue that is visible in these experiments is the discrepancy between classes. While some classes like "Politics" are shared, the topic classification datasets also have language- and location-specific classes like "Nigeria" and "Africa" which are not part of the English fine-tuning dataset. In our experiments, we use the intersection of labels for NER (excluding DATE and MISC for Hausa and Yorùbá) and the union of labels for topic classification.

The results in Figure 1c and in the Appendix confirm the benefits of fine-tuning on high-resource languages already shown in past research. They show, however, also the large gains in performance that can be obtained by training on a minimal number of target instances. While the zero-shot setting in (Hu et al., 2020) is interesting from a methodological perspective, using a small training set for the target language seems much more beneficial for a practical application. In our experiments, we get - with only ten labeled sentences - an improvement of at least 10 points in the F1-score for a shared label set on NER. For topic classification (Figure 1b) the transfer learning is not beneficial, which might be due to the mismatch in the label sets.

6 Distant Supervision

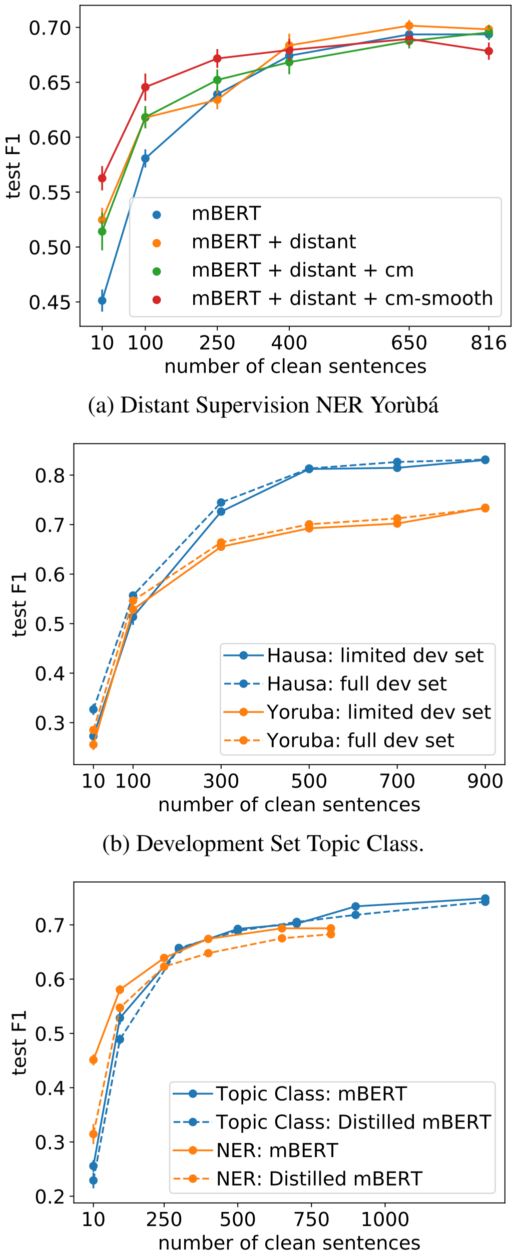
Distant and weak supervision are popular techniques when labeled data is lacking. It allows a domain expert to insert their knowledge without having to label every instance manually. The expert can, e.g. create a set of rules that are then used to label the data automatically (Ratner et al., 2020) or information from an external knowledge source can be used (Rijhwani et al., 2020). This kind of (semi-) automatic supervision tends to contain more errors which can hurt the performance of classifiers (see e.g. (Fang and Cohn, 2016)). To avoid this, it can be combined with label noise handling techniques. This pipeline has been shown to be effective for several NLP tasks (Lange et al., 2019; Paul et al., 2019; Wang et al., 2019; Chen et al., 2019; Mayhew et al., 2019), however, mostly for RNN based approaches. As we have seen in Section 4 that these have a lower baseline performance, we are interested in whether distant supervision is still useful for the better performing transformer models. Several of the past works evaluated their approach only on high-resource languages or simulated low-resource scenarios. We are, therefore, also interested in how the distant supervision performs for the actual resource-lean African languages we study.

To create the distant supervision, native speakers with a background in NLP were asked to write labeling rules. For the NER labels PER, ORG and LOC, we match the tokens against lists of entity names. These were extracted from the corresponding categories from Wikidata. For the DATE label, the insight is used that date expressions are usually preceded by date keywords in Yorùbá, as reported by Adelani et al. (2020). We find similar patterns in Hausa like "*ranar*"(day), "*watan*" (month), and "*shekarar*"(year). For example, *"18th of May, 2019"* in Hausa translates to *"ranar 18 ga watan Mayu, shekarar 2019"*. The annotation rules are based on these keywords and further heuristics. Directly applying this distant supervision on the NER test sets results in an F1-score of 54% and 62% on Hausa and Yorùbá, respectively.

For the topic classification task, the distant supervision rules are based on a dictionary of words relating to each of the classes. To induce the dictionaries, we collected terms related to different classes from web sources. For example, for the "Sport" label, names of sportspeople and sport-related organizations were collected and similarly for the "Africa" label, names of countries, their capitals and major cities and their politicians. To label a news headline, the intersection between each class-dictionary and the text was computed, and a class was selected with a majority voting scheme. We obtain an F1-score of 49% and 55% on the Hausa and Yorùbá test set respectively when applying the distant supervision directly to the topic classification test sets. Additional details on the distant supervision are given in the Appendix.

For label noise handling we use the confusion matrix approach for NER by Hedderich and Klakow (2018), marked as *cm* in the plots. Additionally, we propose to combine it with the smoothing concept by Lv et al. (2020).

The Figures 2a and in the Appendix show that when only a small amount of manually labeled data is available, distant supervision can be a helpful addition. E.g. for the NER task in Yorùbá, combining distant supervision and noise handling with 100 labeled sentences achieves similar performance to using 400 manually labeled sentences. For label noise handling, combining the confusion matrix with the smoothing approach might be beneficial because the estimated confusion matrix is flawed when only small amounts of labeled data are given. When more manually labeled data is available, the noisy annotations lose their benefit and can become harmful to performance. Improved noise-handling techniques might be able to mitigate this.



(a) Distant Supervision NER Yorùbá

(b) Development Set Topic Class.

(c) DistilBERT on Yorùbá

Figure 2: Distant supervision and model variations. Additional plots in the Appendix.

7 Questioning Assumptions

In this section, we want to discuss certain assumptions taken by us and previous work in low-resource learning to see if these hold and what challenges and opportunities they could bring for future work.

7.1 Development Set

Kann et al. (2019) criticized that research on low-resource often assumes the existence of a development set. Addressing this, we perform hyper-parameter optimization on high-resource English data. For early-stopping (to avoid overfitting), the authors experiment with obtaining an early-stop-epoch from the average of several other languages. To avoid this multi-language set-up and the need to obtain labeled data for multiple languages, we suggest using instead a development set downsized by the same factor as the training data. This approach keeps the ratio between training and development set giving the development set a reasonable size to obtain in a low-resource setting. For the setting with ten labeled sentences for training, the same amount was used for the dev set. The results in Figure 2b and in the Appendix show that this has only a small effect on the training performance.

7.2 Hardware Resources

While the multilingual transformer models show impressive improvements over the RNN baselines, they also require more hardware resources. The LSTM-CNN-CRF model, e.g. has ca. 5M parameters compared to mBERT's over 150M parameters. The computing capabilities for training and deploying such models might not always be given in low-resource scenarios. Through personal conversations with researchers from African countries, we found that this can be an issue. There are approaches to reduce model size while keeping a similar performance quality, e.g. the 25% smaller DistilBERT (Sanh et al., 2019). Figure 2c shows that this performs indeed similar in many cases but that there is a significant drop in performance for NER when only few training sentences are available.

7.3 Annotation Time

In (Hu et al., 2020) and (Kann et al., 2020), it is assumed that no labeled training data is available for the target language. In the previous sections, we showed that even with ten labeled target sentences, reasonable model quality can be achieved. For our annotation efforts, we measured on average 1 minute per annotator per sentence for NER and 6 seconds per sentence for topic classification. We, therefore, think that it is reasonable to assume the availability of small amounts of labeled data. Especially, as we would argue that it is beneficial to have a native speaker or language expert involved when developing a model for a specific language.

For distant supervision, a trade-off arises given these annotation times. While extracting named entities from knowledge bases requires minimal manual effort assuming a set-up system, manual crafting rules took 30 minutes for the DATE label and 2.5 hours for each topic classification dataset. When reporting results for distant supervision, the performance benefits should therefore also be compared against manual annotation in the same time frame.

8 Conclusions

In this work, we evaluated transfer learning and distant supervision on multilingual transformer models, studying realistic low-resource settings for African languages. We show that even with a small amount of labeled data, reasonable performance can be achieved. We hope that our new datasets and our reflections on assumptions in low-resource settings help to foster future research in this area.

Acknowledgments

We would like to thank Toyin Aina and Emmanuel Olawuyi for their support in annotating the Hausa data. We also thank Thomas Niesler for his helpful insights as well as Alexander Blatt and the anonymous reviewers for their feedback. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102, the EU-funded Horizon 2020 projects ROXANNE under grant number 833635 and COMPRISE under grant agreement No. 3081705.

References

Idris Abdulmumin and Bashir Shehu Galadanci. 2019. hauwe: Hausa words embedding for natural language processing. *2019 2nd International Conference of the IEEE Nigeria Computer Chapter (NigeriaComputConf)*.

David Ifeoluwa Adelani, Michael A. Hedderich, Dawei Zhu, Esther van den Berg, and Dietrich Klakow. 2020. Distant supervision and noisy label learning for low resource named entity recognition: A study on hausa and yorùbá.

Jesujoba Alabi, Kwabena Amponsah-Kaakyire, David Adelani, and Cristina Espana-Bonet. 2020. Massive vs. Curated Word Embeddings for Low-Resourced Languages. The Case of Yorùbá and Twi. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2747–2755, Marseille, France. European Language Resources Association.

Junfan Chen, Richong Zhang, Yongyi Mao, Hongyu Guo, and Jie Xu. 2019. Uncover the ground-truth relations in distant supervision: A neural expectation-maximization framework. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 326–336, Hong Kong, China. Association for Computational Linguistics.

Artem Chernodub, Oleksiy Oliynyk, Philipp Heidenreich, Alexander Bondarenko, Matthias Hagen, Chris Biemann, and Alexander Panchenko. 2019. Targer: Neural argument mining at your fingertips. In *Proceedings of the 57th Annual Meeting of the Association of Computational Linguistics (ACL'2019)*, Florence, Italy.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734. ACL.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale.

Jacob Devlin. 2019. mBERT README file.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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, Minneapolis, Minnesota. Association for Computational Linguistics.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig (eds.). 2019. Ethnologue: Languages of the world. twenty-second edition.

Roald Eiselen. 2016. Government domain named entity recognition for south African languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3344–3348, Portorož, Slovenia. European Language Resources Association (ELRA).

Roald Eiselen and Martin J. Puttkammer. 2014. Developing text resources for ten south african languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014*, pages 3698–3703. European Language Resources Association (ELRA).

Meng Fang and Trevor Cohn. 2016. Learning when to trust distant supervision: An application to low-resource POS tagging using cross-lingual projection. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 178–186, Berlin, Germany. Association for Computational Linguistics.

Michael A. Hedderich and Dietrich Klakow. 2018. Training a neural network in a low-resource setting on automatically annotated noisy data. In *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP, DeepLo@ACL 2018, Melbourne, Australia, July 19, 2018*, pages 12–18. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization.

Katharina Kann, Kyunghyun Cho, and Samuel R. Bowman. 2019. Towards realistic practices in low-resource natural language processing: The development set. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3342–3349, Hong Kong, China. Association for Computational Linguistics.

Katharina Kann, Ophélie Lacroix, and Anders Søgaard. 2020. Weakly supervised pos taggers perform poorly on truly low-resource languages.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*.

Lukas Lange, Michael A. Hedderich, and Dietrich Klakow. 2019. Feature-dependent confusion matrices for low-resource NER labeling with noisy labels. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3554–3559, Hong Kong, China. Association for Computational Linguistics.

Anne Lauscher, Vinit Ravishankar, Ivan Vulic, and Goran Glavas. 2020. From zero to hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers. *ArXiv*, abs/2005.00633.

Melinda Loubser and Martin J. Puttkammer. 2020a. Viability of neural networks for core technologies for resource-scarce languages. *Information*, 11:41.

Melinda Loubser and Martin J. Puttkammer. 2020b. Viability of neural networks for core technologies for resource-scarce languages. *Information*, 11:41.

Bingfeng Luo, Yansong Feng, Zheng Wang, Zhanxing Zhu, Songfang Huang, Rui Yan, and Dongyan Zhao. 2017. Learning with noise: Enhance distantly supervised relation extraction with dynamic transition matrix. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 430–439, Vancouver, Canada. Association for Computational Linguistics.

Xianbin Lv, Dongxian Wu, and Shu-Tao Xia. 2020. Matrix smoothing: A regularization for DNN with transition matrix under noisy labels. *CoRR*, abs/2003.11904.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

Stephen Mayhew, Snigdha Chaturvedi, Chen-Tse Tsai, and Dan Roth. 2019. Named entity recognition with partially annotated training data. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 645–655, Hong Kong, China. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Debjit Paul, Mittul Singh, Michael A. Hedderich, and Dietrich Klakow. 2019. Handling noisy labels for robustly learning from self-training data for low-resource sequence labeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 29–34, Minneapolis, Minnesota. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Alexander Ratner, Stephen H. Bach, Henry R. Ehrenberg, Jason A. Fries, Sen Wu, and Christopher Ré. 2020. Snorkel: rapid training data creation with weak supervision. *VLDB J.*, 29(2):709–730.

Shruti Rijhwani, Shuyan Zhou, Graham Neubig, and Jaime Carbonell. 2020. Soft gazetteers for low-resource named entity recognition.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Stephanie Strassel and Jennifer Tracey. 2016. LORELEI language packs: Data, tools, and resources for technology development in low resource languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3273–3280, Portorož, Slovenia. European Language Resources Association (ELRA).

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Jennifer Tracey, Stephanie Strassel, Ann Bies, Zhiyi Song, Michael Arrigo, Kira Griffitt, Dana Delgado, Dave Graff, Seth Kulick, Justin Mott, and Neil Kuster. 2019. Corpus building for low resource languages in the DARPA LORELEI program. In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, pages 48–55, Dublin, Ireland. European Association for Machine Translation.

Hao Wang, Bing Liu, Chaozhuo Li, Yan Yang, and Tianrui Li. 2019. Learning with noisy labels for sentence-level sentiment classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6286–6292, Hong Kong, China. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 649–657. Curran Associates, Inc.

A Languages

In this work, we consider three languages: Hausa, isiXhosa and Yorùbá. These languages are from two language families: Niger-Congo and Afro-Asiatic, according to Ethnologue (Eberhard et al., 2019), where the Niger-Congo family has over 20% of the world languages.

The Hausa language is native to the northern part of Nigeria and the southern part of the Republic of Niger with more than 45 million native speakers (Eberhard et al., 2019). It is the second most spoken language in Africa after Swahili. Hausa is a tonal language, but this is not marked in written text. The language is written in a modified Latin alphabet.

Yorùbá, on the other hand, is native to south-western Nigeria and the Republic of Benin. It has over 35 million native speakers (Eberhard et al., 2019) and is the third most spoken language in Africa. Yorùbá is a tonal language with three tones: low, middle and high. These tones are represented by the grave ("\"), optional macron ("−") and acute ("/") accents respectively. The tones are represented in written texts along with a modified Latin alphabet.

Lastly, we consider isiXhosa, a Bantu language that is native to South Africa and also recognized as one of the official languages in South Africa and Zimbabwe. It is spoken by over 8 million native speakers (Eberhard et al., 2019). isiXhosa is a tonal language, but the tones are not marked in written text. The text is written with the Latin alphabet.

Kann et al. (2020) used as an indicator for a low-resource language the availability of data in the Universal Dependency project (Nivre et al., 2020). The languages we study suit their indicator having less than 10k (Yorùbá) or no data (Hausa, isiXhosa) at the time of writing.

B Datasets

B.1 Existing Datasets

The WikiAnn corpus (Pan et al., 2017) provides NER datasets for 282 languages available on Wikipedia. These are, however, only silver-standard annotations and for Hausa and isiXhosa less than 4k and 1k tokens respectively are provided. The LORELEI project announced the release of NER datasets for several African languages via LDC (Strassel and Tracey, 2016; Tracey et al., 2019) but have not yet done so for Hausa and Yorùbá at the time of writing.

Eiselen and Puttkammer (2014) and Eiselen (2016) created NLP datasets for South African languages. We use the latter's NER dataset for isiXhosa. For the Yorùbá NER dataset (Alabi et al., 2020), we use the authors' split into training, dev and test set of the cased version of their data.[2] For the isiXhosa dataset[3], we use an 80%/10%/10% split following the instructions in (Loubser and Puttkammer, 2020b). The split is based on token-count, splitting only after the end of the sentence (information obtained through personal conversation with the authors). For the fine-tuning of the zero- and few-shot models, the standard CoNLL03 NER (Tjong Kim Sang and De Meulder, 2003) and AG News (Zhang et al., 2015) datasets are used with their existing splits.

B.2 New Datasets

B.2.1 Hausa NER

For the Hausa NER annotation, we collected 250 articles from VOA Hausa[4], 50 articles each from the five pre-defined categories of the news website. The categories are Najeriya (Nigeria), Afirka (Africa), Amurka (USA), Sauran Duniya (the rest of the world) and Kiwon Lafiya (Health). We removed articles with less than 50 tokens which results in 188 news articles (over 37K tokens). We asked two volunteers who are native Hausa speakers to annotate the corpus separately. Each volunteer was supervised by someone with experience in NER annotation. Following the named entity annotation in Yorùbá by Alabi et al. (2020), we annotated PER, ORG, LOC and DATE (dates and times) for Hausa. The annotation was based on the MUC-6 Named Entity Task Definition guide.[5] Comparing the annotations of the volunteers, we observe a conflict for 1302 tokens (out of 4838 tokens) excluding the non-entity words (i.e. words with 'O' labels). One of the annotators was better

[2] https://github.com/ajesujoba/YorubaTwi-Embedding/tree/master/Yoruba/Yor%C3%B9b%C3%A1-NER

[3] https://repo.sadilar.org/handle/20.500.12185/312

[4] https://www.voahausa.com

[5] https://cs.nyu.edu/faculty/grishman/NEtask20.book_1.html

in annotating DATE, while the other was better in annotating ORG especially for multi-word expressions of entities. We resolved all the conflicts after discussion with one of the volunteers. The split of annotated data of the Yoruba and Hausa NER data is 70%/10%/20% for training, validation and test sentences.

B.2.2 Hausa and Yorùbá Text classification

For the topic classification datasets, news titles were collected from VOA Hausa and the BBC Yoruba news website[6]. Two native speakers of the language annotated each dataset. We categorized the Yorùbá news headlines into 7 categories, namely "Nigeria", "Africa", "World", "Entertainment", "Health", "Sport", "Politics". Similarly, we annotated 5 (of the 7) categories for Hausa news headlines, excluding "Sport" and "Entertainment" as there was only a limited number of examples. The "Politics" category in the annotation is only for Nigerian political news headlines. Comparing the two annotators, there was a conflict rate of 7.5% for Hausa and 5.8% for Yorùbá. The total number of news titles after resolving conflicts was 2,917 for Hausa and 1,908 for Yorùbá.

C Word Embeddings

For the RNN models, we make use of word features obtained from Word2Vec embeddings for the Hausa language and FastText embeddings for Yorùbá and isiXhosa languages. We utilize the better quality embeddings recently released by Abdulmumin and Galadanci (2019) and Alabi et al. (2020) for Hausa and Yorùbá respectively instead of the pre-trained embeddings by Facebook that were trained on a smaller and lower quality dataset from Wikipedia. For isiXhosa, we did not find any existing word embeddings, therefore, we trained FastText embeddings from data collected from the I'solezwe[7] news website and the *OPUS*[8] parallel translation website. The corpus size for isiXhosa is 1.4M sentences (around 15M tokens). We trained FastText embeddings for isiXhosa using *Gensim*[9] with the following hyper-parameters: embedding size of 300, context window size of 5, minimum word count of 3, number of negative samples ten and number of iterations 10.

D Distant Supervision

D.1 Distant supervision for Personal names, Organisation and Locations

We make use of lists of entities to annotate PER, ORG and LOC automatically. In this paper, we extract personal names, organizations and locations from Wikidata as entity lists and assign a corresponding named entity label if tokens from an unlabeled text match an entry in an entity list.

For NER, we use manual heuristics to improve matching. For Yorùbá, a minimum token length of 3 was set for extraction of LOC and PER, while the minimum length for ORG was set to 2. This reduces the false positive rate, e.g. preventing matches with function words like "of".

Applying this on the test set, we obtained a precision of 80%, 38% and 28% for LOC, ORG and PER respectively; a recall of 73%, 52% and 14% for LOC, ORG and PER respectively; and an F1-score of 76%, 44% and 19% for LOC, ORG and PER respectively.

For Hausa NER, a minimum token length of 4 was set for extraction of LOC, ORG and PER. Based on these manual heuristics, on the test set, we obtained a precision of 67%, 12% and 47% for LOC, ORG and PER respectively; a recall of 63%, 37% and 56% for LOC, ORG and PER respectively; and an F1-score of 65%, 18% and 51% for LOC, ORG and PER respectively.

D.2 DATE rules for NER

Rules allow us to apply the knowledge of domain experts without the manual effort of labeling each instance. We asked native speakers with knowledge of NLP to write DATE rules for Hausa and Yorùbá. In both languages, date expressions are preceded by date keywords, like "*ranar*" / "*ọjọ́*" (day), "*watan*" / "*oṣù*" (month), and "*shekarar*" / "*ọdụ́*n" (year) in Hausa/Yorùbá. For example, *"18th of December, 2019"* in Hausa / Yorùbá translates to " *ranar 18 ga watan Disamba, shekarar 2019*" / "*ọjọ́ 18 oṣù Ọpẹ̀, ọdún 2019*". The annotation rules are based on these three criteria: (1) A token is a date keyword or follows a date keyword in a sequence. (2) A token is a digit, and (3) other heuristics to capture relative dates and date periods connected by conjunctions e.g between July 2019 and March 2020. Applying these rules result in a precision of 49.30%/51.35%, a recall of 60.61%/79.17% and an F1-score of 54.42%/62.30% on Hausa /Yorùbá test set respectively.

[6] https://www.bbc.com/yoruba

[7] https://www.isolezwelesixhosa.co.za/

[8] http://opus.nlpl.eu/

[9] https://radimrehurek.com/gensim/

D.3 Rules for Topic classification

For the Yorùbá topic classification task, we collected terms that correspond to the different classes into sets. For example, the set for the class Nigeria contains names of agencies and organizations, states and cities in Nigeria. The set for the World class is made up of the name of countries of the world, their capitals and major cities and world affairs related organizations. Names of sporting clubs and sportspeople across the world were used for the Sports class and list of artists and actresses and entertainment-related terms for the Entertainment class. Given a news headline to be annotated, we get the union set of 1- and 2-grams and obtain the intersection with the class dictionaries we have. The class with the highest number of intersecting elements is selected. In the case of a tie, we randomly pick a class out the classes with a tie. Just as we did for Yorùbá, we collected the class-related tokens for the Hausa text classification. We, however, split the classification into two steps, checking some important tokens and using the same approach as we used for Yorùbá. If a headline contains the word *cutar* (disease) , it is classified as Health, if it contains tokens such as *inec*, *zaben*, *pdp*,*apc* (which are all politics related tokens) it is classified as Politics. Furthermore, sentences with any of the following tokens *buhari*, *legas*, *kano*, *kaduna*, *sokoto* are classified as Nigeria while sentences with *afurka*, *kamaru* and *nijar* are classified as Africa. If none of the tokens highlighted above is found, we apply the same approach as we did for the Yorùbá setting, which is majority voting of the intersection set of the news headline with a keyword set for each class. Applying these rules results in a precision of $59.54\%/60.05\%$, a recall of $46.04\%/53.66\%$ and an F1-score of $48.52\%/54.93\%$ on the Hausa /Yorùbá test set respectively.

E Experimental Settings

E.1 General

All experiments were repeated ten times with varying random seeds but with the same data (subsets). We report mean F1 test score and standard error ($\sigma/\sqrt{10}$). For NER, the score was computed following the standard CoNLL approach (Tjong Kim Sang and De Meulder, 2003) using the *seqeval* implementation.[10] Labels are in the BIO2-scheme. For evaluating topic classification, the implementation by *scikit-learn* was used.[11] All models are trained for 50 epochs, and the model that performed best on the (possibly size-reduced) development set is used for evaluation.

E.2 BERT and XLM-RoBERTa

As multilingual transformer models, mBert and XLM-RoBERTa are used, both in the implementation by Wolf et al. (2019). The specific model IDs are *bert-base-multilingual-cased* and *xlm-roberta-base*.[12] For the DistilBERT experiment it is *distilbert-base-multilingual-cased*. As is standard, the last layer (language model head) is replaced with a classification layer (either for sequence or token classification). Models were trained with the Adam optimizer and a learning rate of $5e^{-5}$. Gradient clipping of value 1 is applied. The batch size is 32 for NER and 128 for topic classification. For distant supervision and XLM-RoBERTa on the Hausa topic classification data with 100 or more labeled sentences, we observed convergence issues where the trained model would just predict the majority classes. We, therefore, excluded for this task runs where *on the development set* the class-specific F1 score was 0.0 for two or more classes. The experiments were then repeated with a different seed.

E.3 Other Architectures

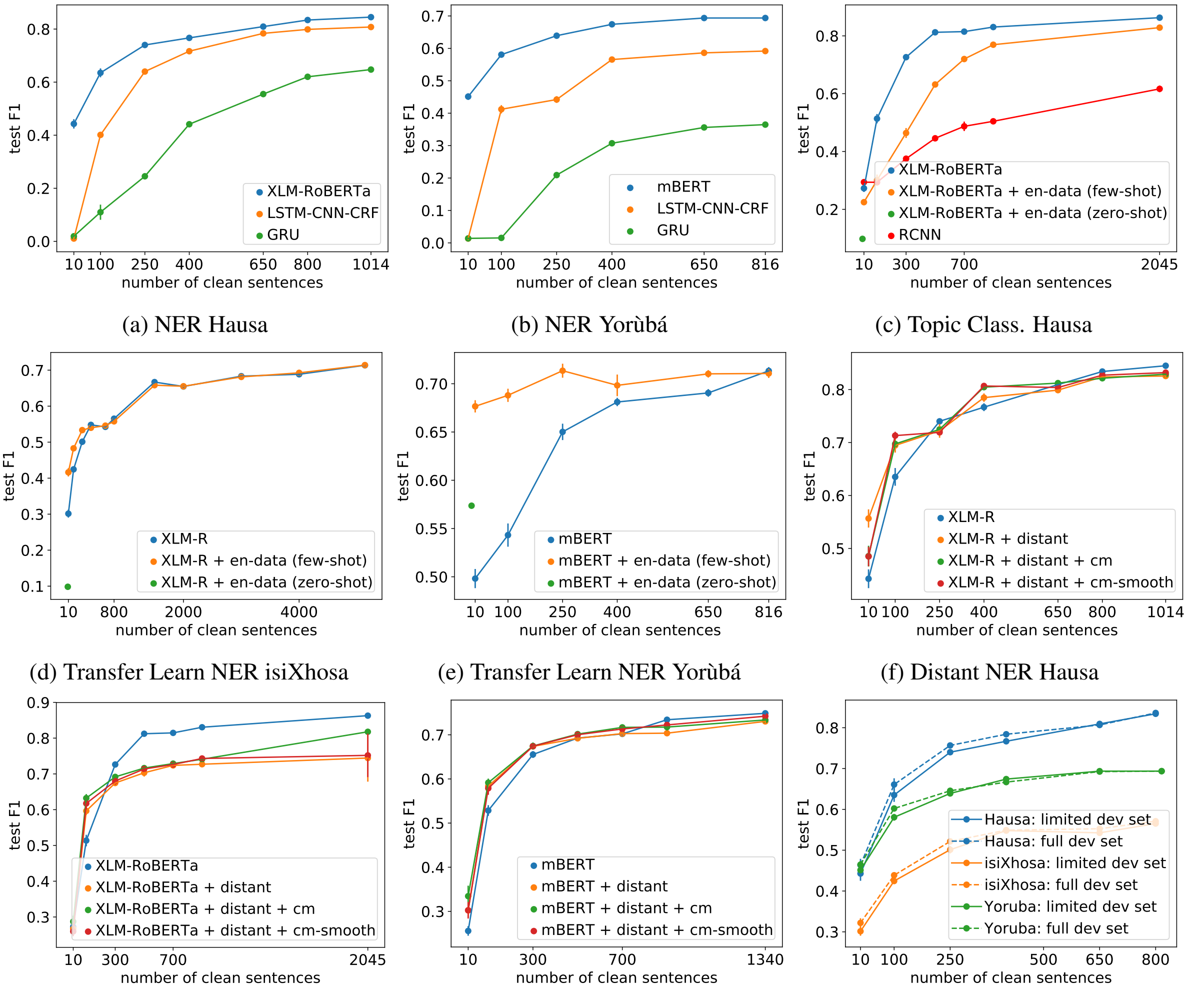
For the GRU and LSTM-CNN-CRF model, we use the implementation by Chernodub et al. (2019) with modifications to support FastText embeddings and the *seqeval* evaluation library. Both model architectures are bidirectional. Dropout of 0.5 is applied. The batch-size is 10 and SGD with a learning rate of 0.01, and a decay of 0.05 and momentum of 0.9 is used. Gradients are clipped with a value of 5. The RNN dimension is 300. For the CNN, the character embedding dimension is 25 with 30 filters and a window-size of 3.

For the topic classification task, we experiment with the RCNN model proposed by (Lai et al., 2015). The hidden size in the Bi-LSTM is 100 for each direction. The linear layer after the Bi-LSTM reduces the dimension to 64. The model is trained for 50 epochs.

[10] https://github.com/chakki-works/seqeval

[11] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html

[12] https://huggingface.co/transformers/pretrained_models.html



(a) NER Hausa

(b) NER Yorùbá

(c) Topic Class. Hausa

(d) Transfer Learn NER isiXhosa

(e) Transfer Learn NER Yorùbá

(f) Distant NER Hausa

(g) Distant Topic Class. Hausa

(h) Distant Topic Class. Yorùbá

(i) Development Set NER

Figure 3: Additional plots.

E.4 Transfer Learning

For transfer learning, the model is first fine-tuned on labeled data from a high-resource language. Following (Hu et al., 2020), we use the English CoNLL03 NER dataset (Tjong Kim Sang and De Meulder, 2003) for NER. It consists of ca. 8k training sentences. The model is trained for 50 epochs and the weights of the best epoch according to the development set are taken. The training parameters are the same as before. On the English CoNLL03 test set, the model achieves a performance of 0.90 F1-score. As the Hausa and Yorùbá datasets have slightly different label sets, we only use their intersection, resulting in the labels PER, LOC and ORG and excluding MISC from CoNLL03 and the DATE label from Hausa/Yorùbá. For isiXhosa, the label sets are identical (i.e. also including MISC). After fine-tuning the model on the high-resource data, the model is directly evaluated on the African test set (for zero-shot) or fine-tuned and then evaluated on the African data (for few-shot).

For topic classification, the AG News corpus is used (Zhang et al., 2015). It consists of 120k training sentences. The model is trained for 20 epochs and the weights of the best epoch according to the test set are used. On this set, an F1-score of 0.93 is achieved. The training procedure is the same as above. For the labels, we use the union of the labels of the AG News corpus (Sports, World, Business and Sci/Tech) and the African datasets.

E.5 Label Noise Handling

We use a confusion matrix which is a common approach for handling noisy labels (see, e.g. (Fang and Cohn, 2016; Luo et al., 2017; Lange et al., 2019; Wang et al., 2019)). The confusion matrix models the relationship between the true, clean label of an instance and its corresponding noisy label. When training on noisy instances, the confusion matrix is added to the output of the main model (that usually predicts clean labels) changing the output label distribution from the clean to the noisy one. This allows to then train on noisily labeled instances without a detrimental loss obtained by predicting the true, clean label but having noisy, incorrect labels as targets.

We use the specific approach by Hedderich and Klakow (2018) that was developed to work with small amounts of manually labeled, clean data and a large amount of automatically annotated, noisy labels obtained through distant supervision. To get the confusion matrix of the noise, the distant supervision is applied on the small set of clean training instances. From the resulting pairs of clean and noisy labels, the confusion matrix can be estimated.

In a setting where only a few instances are available, the estimated confusion matrix might not be close to the actual change in the noise distribution. We, therefore, combine it with the smoothing approach by Lv et al. (2020). Each entry of the probabilistic confusion matrix is raised to the power of β and then row-wise normalized.

As studied by Hedderich and Klakow (2018), we do not use the full amount of available, distantly supervised instances in each epoch. Instead, in each epoch, only a randomly selected subset of the size of the clean, manually labeled training data is used to lessen the negative effects of the noisy labels additionally. For smoothing, $\beta = 0.8$ is used as this performed best for Lv et al. (2020).