The Impact of Data Corruption on Named Entity Recognition for Low Resourced Languages

Abstract

Data is often a massively limiting factor in natural language processing for low-resourced languages. In particular, there is significantly less data than for higher-resourced languages. This data is also often of low-quality, rife with errors and invalid text. Many prior works focus on dealing with these problems, either by generating synthetic data, or filtering out low-quality parts thereof. We instead investigate these factors more deeply, by systematically measuring the effect of data quantity and quality on the performance of pre-trained language models. Our results show that having a missing annotation is preferred compared to having an incorrect one; and that models can perform remarkably well with only 10% of the training data. Finally, all of our results are very consistent across 11 languages and 4 different pre-trained models.

1 Introduction

Natural Language Processing (NLP) is an impactful field, and has received much interest recently, and has been applied in numerous settings [Vaswani *et al.*, 2017; Conneau *et al.*, 2020]. However, much of the focus is on high-resourced languages [Vaswani *et al.*, 2017; Conneau *et al.*, 2020; Radford *et al.*, 2018, 2019], such as English, German, Spanish, etc. While this has led to impressive results for these languages, lower-resourced languages have often not enjoyed as much attention, leading to a large gap in NLP system performance between high- and low-resourced languages. This gap has resulted in an increasingly large body of work focused exclusively on low-resourced languages, either developing models [Ogueji *et al.*, 2021; Alabi *et al.*, 2022] or introducing datasets [Oyewusi *et al.*, 2021; Adelani *et al.*, 2021, 2022a,b].

Despite this recent work and impressive progress, data is still one large limiting factor for low-resourced NLP [Adelani *et al.*, 2022a,b]. In particular, the two main problems are quality and quantity of data. First of all, the amount of data available for low-resourced languages is often a fraction of the high-resourced languages; and for many languages, no data exists at all. Secondly, the data that is available often has questionable quality, containing corruption, invalid tokens or just gibberish [Kreutzer *et al.*, 2022], which has detrimental downstream effects on the models trained using this data [Abdul-Rauf *et al.*, 2012; Alabi *et al.*, 2019].

The large amounts of poor-quality data has led to many works focusing on filtering data [Axelrod *et al.*, 2011; Xu *et al.*, 2019; Imankulova *et al.*, 2017; Abdulmumin *et al.*, 2021, 2022] to improve results by discarding all of the invalid or corrupt portions of a dataset. In many of these works, the perspective is that there is an existing, but noisy, dataset, and it must be filtered, keeping only the high-quality parts thereof. The data quality is often so poor that having a smaller high-quality dataset can be better than having a much larger, and lower-quality, dataset. The lack of data has also spurred research into synthetically generating more data, by using techniques such as backtranslation [Bojar and Tamchyna, 2011; Lambert *et al.*, 2011; Sennrich *et al.*, 2016] or using a translation model to generate labelled data for one language using by translating another language's dataset.

We instead take a different perspective, and focus on analysing the effect of systematically reducing the quality of datasets, and examine the implications of this. In particular, we focus on a Named-Entity Recognition (NER) task due to its prevalence in many NLP systems and the existence of a few high-quality datasets in low-resourced languages. We further focus on fine-tuning existing pre-trained language models, as this is a common and high-performing approach, especially for low-resourced languages [Ogueji *et al.*, 2021; Adelani *et al.*, 2021; Alabi *et al.*, 2022]. We specifically focus on corrupting the training datasets in specific ways to examine the effects of data quality on the performance of models. We further alter the amount of data that the models train on to investigate the effect of the amount of training data on performance.

Our results reveal an interesting phenomenon, where the performance dropoff is not linear, i.e. training on 10% of the data does not result in 10% of the performance of training on the entire dataset, but rather close to 80%. We further find that having a wrong label in NER is more damaging than having a missing label – suggesting that when annotators are uncertain, leaving out an annotation is preferable compared to having a wrong one. Finally, we note that our results are consistent across 11 languages and 4 pre-trained models, suggesting that these observations are generally valid.

2 Background and Related Work

2.1 Named Entity Recognition

Named entity recognition (NER) is a token classification task, where the task is to classify each token in a text as an Organisation, Location, Person, Date, or "Other". NER as a field has many impactful applications [Sang and Meulder, 2003; Lample and Chaplot, 2017].

A typical NER dataset consists of multiple sentences, with each sentence containing both the words and their associated labels. The prevailing approach to train NER models is to use a pre-trained large language model (such as BERT [Devlin *et al.*, 2019], XLM-Roberta [Conneau *et al.*, 2020], etc.) and fine-tune it on a small amount of NER data [Conneau *et al.*, 2020; Adelani *et al.*, 2021].

2.2 Data Collection and Annotation

Since the lack of data has traditionally been a major limiting factor for low-resourced NLP research, multiple different approaches have developed to effectively collect data in resource-constrained settings. In particular, community involvement has played a large part in this [Nekoto *et al.*, 2020, 2022], where native speakers annotate or create datasets to be used in research. This has led to the creation of many different datasets [Adelani *et al.*, 2021; Nekoto *et al.*, 2022], but it relies on community members instead of trained annotators, which may result in some aspects of the annotation being less accurate. Furthermore, while this approach can successfully develop datasets for low-resourced languages, due to logistic challenges and a limited amount of unlabelled text, these datasets are often significantly smaller than high-resourced datasets [Conneau *et al.*, 2020; Adelani *et al.*, 2021].

2.3 Lack of Quality and Quantity in Low-resourced Languages

While there has been significantly progress in recent years, datasets low-resource language are often quite small and limited, or exhibit subpar quality. Both of these factors can lead to badly-performing models. For instance, Kreutzer *et al.* [2022] perform a large-scale audit of several web-scale and automatically extracted multilingual datasets, and find that the quality is often poor, with rubbish characters and sentences being commonplace. Other work has focused instead on the effect of quality on the performance of downstream models.

This lack of quality can have great effects. Alabi *et al.* [2019] show that for certain low-resourced African languages, using a significantly smaller, but curated dataset outperforms training a model on a large, but noisy dataset. Abdulmumin *et al.* [2022] find similar results, where training on filtered data of higher quality improved the performance of translation models for low-resourced languages.

These two observations; that data is often of low-quality and that a smaller, higher-quality dataset is often preferred has led to much work being done on filtering existing datasets, to extract a smaller, but higher-quality subset of sentences. For instance, Abdulmumin *et al.* [2022] filter a large, automatically-aligned dataset, and find that this improved the performance of translation models for low-resourced languages.

The second factor that can limit progress in NLP is a lack of datasets for many languages, or much smaller datasets for low-resourced languages compared to higher-resourced ones [Adelani *et al.*, 2022a].

3 Methodology

Our aim is to analyse and quantify the impact of data corruption on the performance of pre-trained language models. It would allow us to better reason about the importance of quality and quantity of data, ideally informing the future data creation processes for low-resourced languages.

While we can corrupt NER text corpora in various ways, we choose corruptions that simulate a mislabelling scenario during the annotation process, e.g. mislabelling a person in a sentence as an organisation. There are two main reasons for this choice. Firstly, many NER datasets are formed by taking an existing text source, which is usually of high quality, such as news data [Adelani *et al.*, 2021] and annotating each word; thus, errors are more likely to crop up during the annotation process. Secondly, it is challenging to corrupt the base sentences in a reasonable, quantifiable and incremental way, as sentences encompass meaning which is often hard to change atomically.

Thus, we focus on corrupting only the labels, using different strategies detailed in Section 3.1. For each corruption strategy, we uniformly vary the amount of corruption and train our models using the new, corrupted dataset. This process allows us to evaluate how each corruption strategy affects the model as we vary the degree of corruption. We discussed the data for this study in Section 3.2.

3.1 Different Corruption Strategies

We describe the corruption strategies in the next sections, and Figure 1 shows some examples of the different strategies we consider. For corruption strategies involving the quality of labels, we consider the atomic element to be a single NER entity label, even if this consists of multiple words. Thus, when we change the label of a specific entity, we always change its span instead of just a part thereof. Also, we only change the train data while leaving the evaluation data unchanged. The reason for this is because we want an objective comparison of different corruption strategies. A visual illustration of our quality-related corruption strategies can be found in Figure 1.

Sentence Pruning

Dataset annotation is generally an expensive and logistically challenging process (ToDo cite this) when more than one participants are involved. As a result, low-resourced NLP datasets are often not particularly big. Due to this observation, we first evaluate the effect of varying the amount of data available to the models. In this strategy, we randomly remove sentences from the original dataset to create sub-datasets with fewer sentences than the original dataset. This process allows us to measure the model performance as a function of data quantity. We choose to represent quantity as a function of the number of sentences because removing words can alter the

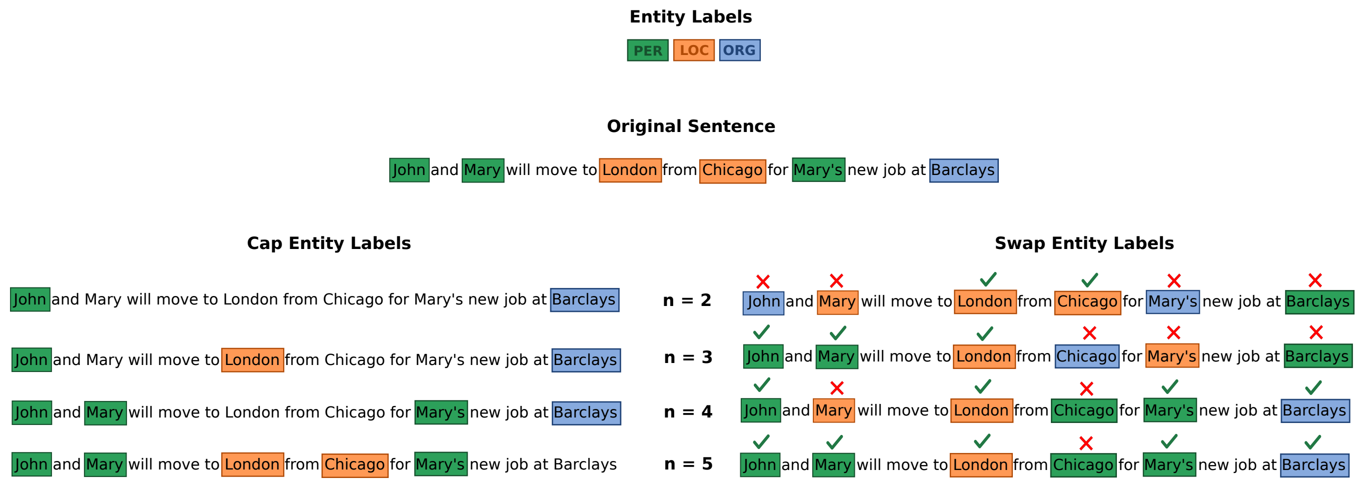

Figure 1: An illustration of the different corruption strategies we use. (Left) When capping labels, we effectively remove a certain number of labels, replacing them with "O". (Right) When swapping labels, we instead randomly replace a label with an incorrect one. In these figures we illustrate the *local* version of the corruptions, with n being the parameter that determines the number of labels kept unchanged.

meaning impacting the model performance in ways we can't control.

Entity Label Capping

A rich NER dataset would be a dataset that has a high annotation density, i.e. a high number of annotated entities per sentence. This strategy aims at inhibiting the model by thresholding the number of entity annotations allowed in the dataset. In the real world, this would be equivalent to a situation where an annotator failed to label a particular span of token as one of entities PER, LOC, ORG, DATE, instead giving it the default entity type O, which generally means *not relevant*.

We have two variations of this corruption, *local* and *global*. Local means that we have a per-sentence threshold for the number of annotations; for instance, if this threshold is 2, we keep only 2 annotations per sentence, deleting (i.e. setting to O) the others. Global, on the other hand, means that we keep only a certain percentage of labels across the entire dataset; for example, 50% would mean that we randomly remove half of all the labels. This, in contrast to the local setting, may leave some sentences unmodified, or completely remove all annotations from certain sentences.

Entity Label Swapping

Another scenario that could happen during the annotation procedure would be the mislabelling of a span of tokens with the wrong entity. For example, the organisation *John Deere* is mistakenly labelled as a person. This creates a situation where there are contradictory labels, with the same token potentially having different labels, some correct and some incorrect. The goal behind this strategy is to determine how robust large pre-trained language models are to such contradictions. In the same spirit as the previous corruption strategies, we corrupt the datasets at a local and global level by either setting a threshold per sentence or across the entire corpus.

3.2 Data

We use the MasakhaNER dataset [Adelani *et al.*, 2021], which is a high-quality dataset for 10 low-resourced, African languages. We specifically focus on low-resourced languages, as these languages often suffer from the aforementioned quality and quantity problems. Furthermore, this dataset is of high-quality, which allows us to evaluate the full spectrum of quality, from gold-standard to completely corrupted. Additionally, we have 10 different languages to evaluate how much the specific language affects the results.

Table 1: The number of sentences for each NER dataset we consider.

Language	Number of Sentences
hau	1912
pcm	2124
ibo	2235
lug	1428
kin	2116
wol	1871
conll_2003_en	14042
luo	644
swa	2109
amh	1750
yor	2171

Finally, as a baseline, we also use the CONLL NER dataset, which is a staple NER dataset in English [Sang and Meulder, 2003].

Table 1 contains information about the relative sizes of each language's dataset and Figure 2 shows how many entities of each type there are per language.

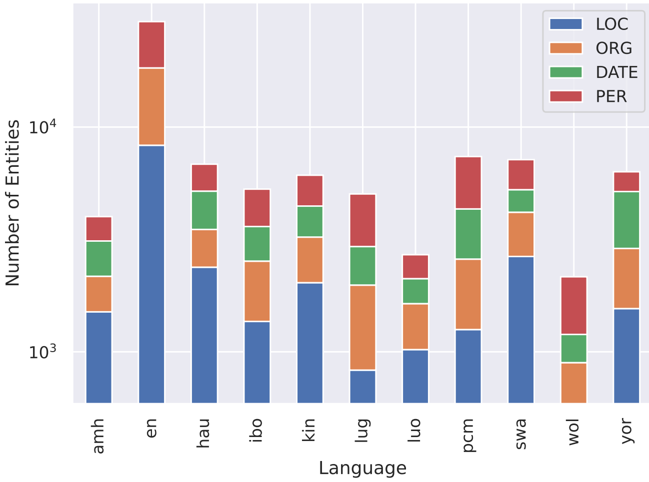

Figure 2: Number of Entities per Language

3.3 Metrics

In addition to considering the overall classification F1 score for each model, we consider the uncertainty of the models as the level of data corruption increases. This allows us quantify the effect of corruption on the certainty and confidence of the models. In particular, we use the *entropy* **TODO**.

4 Experiments

Having described our corruption strategies, we now perform our experiments and showcase our results. We consider the five corruption strategies described above, and use 4 different pre-trained language models, described in Section 4.1. Each run consists of fine-tuning a single pre-trained model on a single language's dataset, either the original one or a corrupted version. We run all experiments over 3 seeds and average the results to obtain a more accurate performance estimation. The metric we use is the F1 score, as that is commonly used as the main metric when evaluating NER models [Sang and Meulder, 2003; Adelani *et al.*, 2021]. We specifically investigate the effect of progressively corrupting data on the performance of each model. This simulates the effect of having subpar data quality (for instance due to incorrect annotations), but allows us to study this in a controlled setting. We do not modify the test datasets at all.

Then, to normalise results across languages, we divide each F1 score by the value obtained when training with the full, uncorrupted dataset. This effectively measures what fraction of performance is lost and allows us to transform all of the metrics to fall between 0 and 1, resulting in the results being comparable across languages and models. We then average all of our results over the 11 languages and plot the mean and standard deviation. We first examine the effect of quantity in Section 4.3 and then move to understanding the implications of data quality in Section 4.4.

4.1 Different Pre-trained Language Models

We use four different pre-trained language models. We first consider two models developed specifically for low-resourced African languages, AfriBERTa and Afro-XLM-R. AfriBERTa [Ogueji *et al.*, 2021] was pre-trained on less than 1GB of African language text. Afro-XLM-R [Alabi *et al.*, 2022] used *language adaptive fine-tuning*, where a pre-trained language model is fine-tuned on unlabelled data using the same objective as was used during pre-training. Afro-XLM-R performed this process on 20 languages, 17 of them from Africa, starting from XLM Roberta. XLM Roberta [Conneau *et al.*, 2020] is a high-performing model that was pre-trained on 100 languages. Finally, Multilingual BERT [Devlin *et al.*, 2019] used the standard BERT training process on 104 languages using data from Wikipedia.

We have two models pre-trained or adaptively fine-tuned on low-resourced, African languages, some of which are contained in our dataset. The other two models are traditional multilingual models, with the majority of the training datasets consisting of high-resourced languages. All of these models have been shown to perform well in the NER task [Adelani *et al.*, 2021; Ogueji *et al.*, 2021; Alabi *et al.*, 2022]. We choose the specific model versions to be roughly comparable in terms of number of parameters. More information about the models is included in Table 2.

4.2 Initial Results

In Table 3, we show the results when training on the entire training dataset, grouped by model. Overall, most models perform well on most languages, with Afro-XLM-R performing the best on average. mBERT, on the other hand, performs the worst overall, and even has 0 F1 score on Amharic, as it was not pre-trained on data containing this script [Adelani *et al.*, 2021].

4.3 Quantity

Here we investigate the effect of data quantity on the performance of the models. We specifically remove a certain percentage of the data and train the model on the remaining ones. Furthermore, since the specific fraction we keep/delete may have an effect, we run this experiment three times, each time with different random selections of data. We average over these three permutations, and find that the results are very similar across them.

In Figure 3, we examine the performance when randomly removing a certain percentage of sentences from the datasets. The results here emphasise that the relationship between quantity and performance is highly non-linear. For instance, when only having 60% of the sentences from the original dataset, the performance is nearly the same as using 100% of the data. Even more shockingly, when keeping only 10% of the data, AfriBERTa still retains 80% of the performance of training on the full data. Other models, such as XLMR, perform slightly worse, but still reaches roughly 70% performance at 10% of the data. When using even less data, such as 1% or 5%, we do see a sharp dropoff and much worse performance. This may be due to the phenomenon observed by Mindermann *et al.* [2022], where there are many datapoints that do not add in additional information, and do not need to be relearnt.

Table 2: Information about the different pre-trained language models we use.

Name	Name	Source	Parameters	African Languages
AfriBERTa	`afriberta-large`	Ogueji *et al.* [2021]	126M	amh, hau, ibo, kin, pcm, swa, yor
Afro-XLM-R	`afro-xlmr-base`	Alabi *et al.* [2022]	270M?	amh, hau, ibo, kin, pcm, swa, yor
XLM Roberta	`xlm-roberta-base`	Conneau *et al.* [2020]	270M?	amh, hau, swa
Multilingual BERT	`bert-base-multilingual-cased`	Devlin *et al.* [2019]	110M	swa, yor

Table 3: The performance of each pre-trained model when fine-tuning on unaltered training data. The best performance per language is marked in bold.

Model Language	AfriBERTa	Afro-XLM-R	XLM-R	mBERT	Average
amh	0.72 (0.01)	**0.76 (0.02)**	0.72 (0.01)	0.00 (0.0)	0.55
en	0.89 (0.0)	**0.93 (0.0)**	0.93 (0.0)	0.93 (0.0)	0.92
hau	0.90 (0.0)	**0.91 (0.0)**	0.90 (0.0)	0.87 (0.01)	0.89
ibo	**0.87 (0.0)**	0.87 (0.01)	0.83 (0.0)	0.85 (0.0)	0.85
kin	0.74 (0.01)	**0.78 (0.0)**	0.72 (0.01)	0.71 (0.01)	0.74
lug	0.79 (0.0)	**0.81 (0.0)**	0.78 (0.0)	0.80 (0.01)	0.79
luo	0.68 (0.01)	0.69 (0.05)	0.69 (0.02)	**0.72 (0.01)**	0.70
pcm	0.86 (0.01)	**0.89 (0.0)**	0.86 (0.01)	0.88 (0.0)	0.87
swa	0.88 (0.01)	**0.88 (0.0)**	0.87 (0.01)	0.86 (0.01)	0.87
wol	0.61 (0.01)	**0.66 (0.02)**	0.64 (0.01)	0.63 (0.01)	0.64
yor	0.79 (0.01)	**0.81 (0.01)**	0.77 (0.01)	0.79 (0.01)	0.79
Average	0.79	0.82	0.79	0.73	0.78

4.4 Quality

We now consider the effect of quality on performance, specifically looking at the different corruption strategies mentioned in Section 3.

4.5 Global

In Figure 4, we delete a specific fraction of labels, replacing them with the wrong value "O". Here, the performance dropoff is not quite linear, with keeping only 60% of the labels resulting in 80% performance. Even having only 40% of the labels provides 60% performance.

Figure 5 considers a similar scenario, but instead swaps a fraction of the labels with another incorrect entity label. Here we see a much more linear relationship, indicating that performance is more sensitive to the wrong label than a missing one.

4.6 Local

Now, considering the local perturbations, Figure 6 caps the number of labels per sentence, and replaces all excess labels with "O". Here we see that having two labels per sentence is sufficient to recover 80% performance, and 4 entities is enough to obtain near full performance. In Figure 7, we see a more aggressive trend, where swapping one label per sentence with an incorrect one results in 60% performance, and swapping 2 results in an abysmal 30%. This again emphasises that incorrect labels are far more damaging to the model's performance compared to merely removing labels.

4.7 Entropy

In this section we consider the effects of our data corruptions on the entropy of the models. When we progressively delete more sentences in Figure 8, the entropy steadily increases. In Figure 9, we see that when we delete labels randomly, the entropy increases, reaches its peak around keeping 70% of the labels, and then decreases again. When swapping labels, in Figure 10, increasing the fraction of swapped labels decreases the entropy.

5 Discussion & Future Work

Our results shown above are interesting, and highlight quite a few important points. First of all, in most cases, all models exhibit roughly equal behaviour, in terms of the percentage dropoff in performance as the level of corruption increases. This ranges from the Africa-centric models (AfriBERTa and Afro-XLM-R) to the predominantly high-resourced models (XLMR and mBERT). This suggests that the behaviour we see here is quite general, as opposed to being particularly model-specific. The variation across languages is also remarkably low.

Secondly, we find that the quantity of data one trains on does not play a massive role in the final performance obtained; we could get around 80% performance with 10-20% of the sentences of the original dataset. This suggests that we do not need a lot of data to perform well, supporting prior findings [Adelani *et al.*, 2022a]. This means that even modest data-collection and annotation efforts should be able to result in datasets that are large enough to obtain decent performance. Our results are consistent across both the languages contained in the pre-trained models' datasets and those that were not.

Thirdly, the type of corruption can have a large effect on the final performance. For instance, merely leaving out annotations (e.g. replacing them with "O") can still result in high-performance; e.g. when keeping only 60% of the labels, we still obtain 80% performance. On the other hand, when we swap labels with incorrect ones, the relationship is much more linear. This suggests that having incorrect annotations is more harmful than having no annotations at all – which could help inform annotator training in data collection endeavours.

Finally, we find that the density of entities can be of great importance. For instance, deleting 60% of all sentences results in higher performance than deleting 60% of all entity labels and keeping all of the sentences. Thus, having fewer sentences could be feasible, provided each of the sentences is completely, and accurately, annotated.

There are numerous avenues for future work. One option would be to expand our work into other NLP tasks; for example, additional token classification tasks such as parts of speech tagging, or tasks such as machine translation. Additionally, developing more corruption strategies that cover other components of quality would be promising as well. Furthermore, combining multiple different corruption strategies and investigating how robust models are to these. Finally, us-

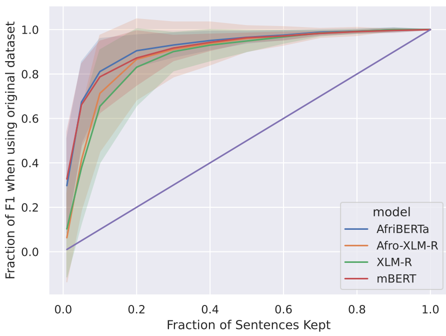

Figure 3: Showing the effect of training on a subset of data on the final performance of the models.

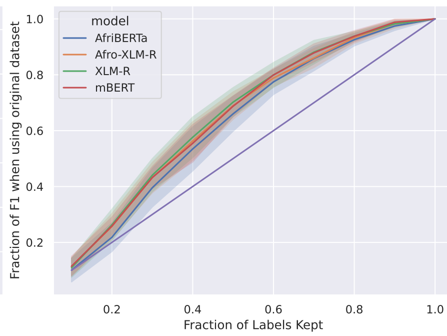

Figure 4: Showing the effect of deleting a certain fraction of labels across the entire dataset, replacing them with O.

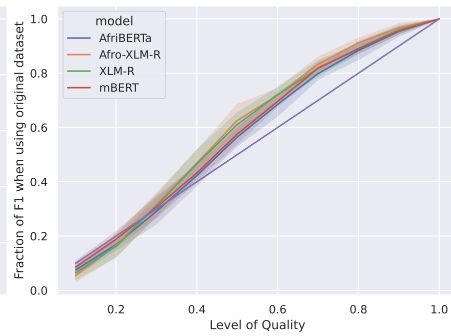

Figure 5: Showing the effect of swapping a certain fraction of labels across the entire dataset, replacing the correct annotation with an incorrect entity.

Here we plot the performance (measured as a fraction of "optimal" performance – where the entire dataset is used) as we change the (a) fraction of the sentences we use for training or (b) the level of corruption in a dataset. These three figures contain the *global* corruptions, where we alter the entire dataset according to some corruption percentage. Standard Deviation across 11 languages is shaded.

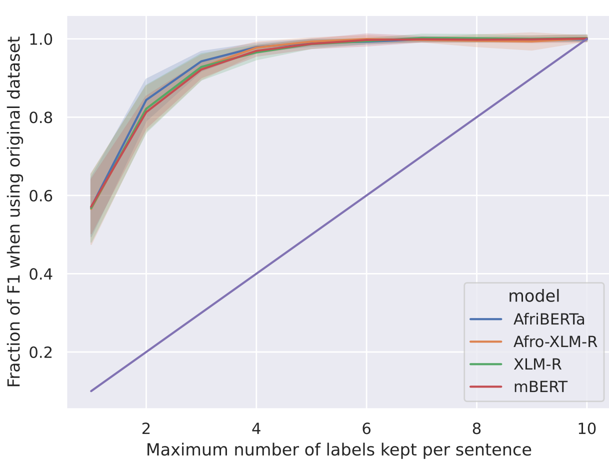

Figure 6: Local Cap Labels

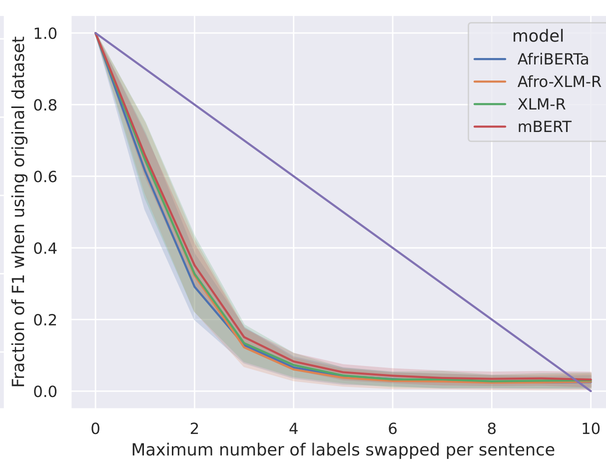

Figure 7: Local Swap Labels

Here we plot the local corruptions strategies, where we either (a) cap the number of labels per sentence, replacing the excess ones with O; or (b) swap a certain number of labels per sentence with an incorrect entity. Standard Deviation across 11 languages is shaded.

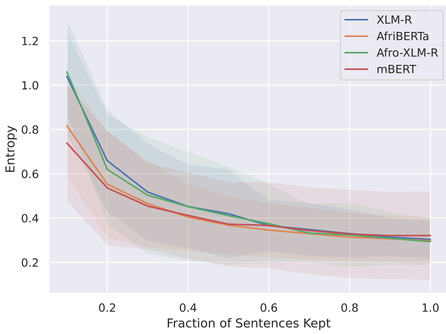

Figure 8: Showing the effect of training on a subset of data on the final performance of the models.

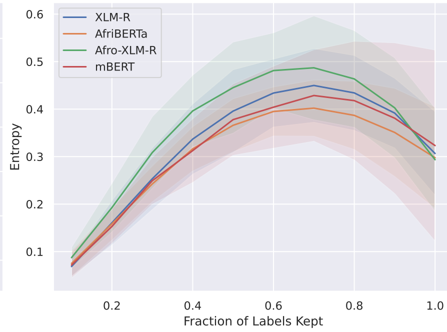

Figure 9: Showing the effect of deleting a certain fraction of labels across the entire dataset, replacing them with O.

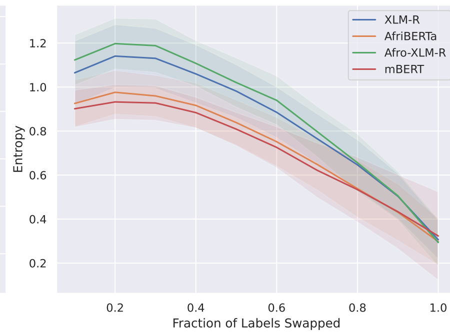

Figure 10: Showing the effect of swapping a certain fraction of labels across the entire dataset, replacing the correct annotation with an incorrect entity.

Here we show the effect of each data corruption strategy on the token entropy of the models.

ing our analysis, future work could develop better methods for dealing with corrupted datasets that would mitigate some of the effect of training on subpar data.

6 Conclusion

Our main aim in this paper is to systematically analyse the effect of data quality and quantity on the performance of pre-trained models on an NER task for low-resourced languages. We do this by designing multiple corruption strategies, and fine-tuning models on various degrees of corruption. Overall, our results emphasise that pre-trained models can perform quite well with remarkably little data, and that missing annotations are less harmful than misleading ones. Ultimately, we hope that this analysis can help inform future NER dataset creation endeavours, or help NLP practitioners when needing to decide on a dataset.

References

Sadaf Abdul-Rauf, Mark Fishel, Patrik Lambert, Sandra Noubours, and Rico Sennrich. Extrinsic evaluation of sentence alignment systems. 2012.

Idris Abdulmumin, Bashir Shehu Galadanci, Abubakar Isa, Habeebah Adamu Kakudi, and Ismaila Idris Sinan. A Hybrid Approach for Improved Low Resource Neural Machine Translation using Monolingual Data. *Engineering Letters*, 29(4):339–350, 2021.

Idris Abdulmumin, Michael Beukman, Jesujoba O. Alabi, Chris Emezue, Everlyn Asiko, Tosin Adewumi, Shamsuddeen Hassan Muhammad, Mofetoluwa Adeyemi, Oreen Yousuf, Sahib Singh, and Tajuddeen Rabiu Gwadabe. Separating grains from the chaff: Using data filtering to improve multilingual translation for low-resourced african languages. In *Proceedings of the Seventh Conference on Machine Translation: Shared Task Papers*, 2022.

David Ifeoluwa Adelani, Jade Z. Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen Hassan Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin P. Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane Mboup, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima Diop, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. Masakhaner: Named entity recognition for african languages. *Trans. Assoc. Comput. Linguistics*, 9:1116–1131, 2021.

David Ifeoluwa Adelani, Jesujoba O. Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Hassan Muhammad, Guyo Dub Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Wambui Gitau, Jade Z. Abbott, Mohamed Ahmed, Millicent Ochieng, Aremu Anuoluwapo, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. A few thousand translations go a long way! leveraging pre-trained models for african news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 3053–3070. Association for Computational Linguistics, 2022.

David Ifeoluwa Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba O. Alabi, Shamsuddeen Hassan 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 Ouoba Kabore, Chris Chinenye Emezue, Aremu Anuoluwapo, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Elvis Mboning, Tajuddeen Gwadabe, Tosin P. Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo L. Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Adeyemi, Gilles Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu Ngoli, and Dietrich Klakow. Masakhaner 2.0: Africa-centric transfer learning for named entity recognition. *CoRR*, abs/2210.12391, 2022.

Jesujoba O. Alabi, Kwabena Amponsah-Kaakyire, David Ifeoluwa Adelani, and Cristina España-Bonet. Massive vs. curated word embeddings for low-resourced languages. the case of yorùbá and twi. *CoRR*, abs/1912.02481, 2019.

Jesujoba O Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. Multilingual language model adaptive fine-tuning: A study on african languages. *CoRR*, abs/2204.06487, 2022.

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 355–362. ACL, 2011.

Ondrej Bojar and Ales Tamchyna. Improving translation model by monolingual data. In Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan, editors, *Proceedings of the Sixth Workshop on Statistical Machine Translation, WMT@EMNLP 2011, Edinburgh, Scotland, UK, July 30-31, 2011*, pages 330–336. Association for Computational Linguistics, 2011.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

Aizhan Imankulova, Takayuki Sato, and Mamoru Komachi. Improving low-resource neural machine translation with filtered pseudo-parallel corpus. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 70–78, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Javier Ortiz Suárez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Balli, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. Quality at a glance: An audit of web-crawled multilingual datasets. *Trans. Assoc. Comput. Linguistics*, 10:50–72, 2022.

Patrik Lambert, Holger Schwenk, Christophe Servan, and Sadaf Abdul-Rauf. Investigations on translation model adaptation using monolingual data. In Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan, editors, *Proceedings of the Sixth Workshop on Statistical Machine Translation, WMT@EMNLP 2011, Edinburgh, Scotland, UK, July 30-31, 2011*, pages 284–293. Association for Computational Linguistics, 2011.

Guillaume Lample and Devendra Singh Chaplot. Playing FPS games with deep reinforcement learning. In Satinder Singh and Shaul Markovitch, editors, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 2140–2146. AAAI Press, 2017.

Sören Mindermann, Jan Markus Brauner, Muhammed Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N. Gomez, Adrien Morisot, Sebastian Farquhar, and Yarin Gal. Prioritized training on points that are learnable, worth learning, and not yet learnt. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *Interna-*

tional Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 15630–15649. PMLR, 2022.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi E. Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Hassan Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Z. Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkabir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Espoir Murhabazi, Elan Van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Itoro Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. Participatory research for low-resourced machine translation: A case study in african languages. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 2144–2160. Association for Computational Linguistics, 2020.

Wilhelmina Nekoto, Julia Kreutzer, Jenalea Rajab, Millicent Ochieng, and Jade Abbott. Participatory translations of oshiwambo: Towards sustainable culture preservation with language technology. In *3rd Workshop on African Natural Language Processing*, 2022.

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*, pages 116–126, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

Wuraola Fisayo Oyewusi, Olubayo Adekanmbi, Ifeoma Okoh, Vitus Onuigwe, Mary Idera Salami, Opeyemi Osakuade, Sharon Ibejih, and Usman Abdullahi Musa. Naijaner : Comprehensive named entity recognition for 5 nigerian languages. *CoRR*, abs/2105.00810, 2021.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In Walter Daelemans and Miles Osborne, editors, *Proceedings of the Seventh Conference on Natural Language Learning, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada, May 31 - June 1, 2003*, pages 142–147. ACL, 2003.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics, 2016.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.

Guanghao Xu, Youngjoong Ko, and Jungyun Seo. Improving neural machine translation by filtering synthetic parallel data. *Entropy*, 21(12):1213, 2019.