ANALYSING THE EFFECTS OF TRANSFER LEARNING ON LOW-RESOURCED NAMED ENTITY RECOGNITION PERFORMANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Transfer learning has led to large gains in performance for nearly all NLP tasks while making downstream models easier and faster to train. This has also been extended to low-resourced languages, with some success. We investigate the properties of transfer learning between 10 low-resourced languages, from the perspective of a named entity recognition task, specifically how much adaptive fine-tuning improves performance, the efficacy of zero-shot transfer as well as the effect of learning on the contextual embeddings computed from the model. Our results give some insight into zero-shot performance as well as the impact of different training schemes and data overlap between the training and testing languages. Particularly, we find that models with the best generalisation to other languages suffer in individual language performance, while models that perform well on a single language often do so at the expense of generalising to others. In the interest of reproducibility, we release an anonymised version of the code[1], and the full version, including models, will be publicly released after the review process is complete.

1 INTRODUCTION

The technique of using a pre-trained Natural Language Processing (NLP) model and fine-tuning it on task-specific data has recently taken the NLP world by storm, achieving state of the art scores in many different tasks (Jiang et al., 2020; Hendrycks et al., 2021; Raffel et al., 2020). Although much of the focus of pre-trained models is on English (Devlin et al., 2019; Radford et al., 2018; 2019), there are also monolingual models for different languages (de Vries et al., 2019; Masala et al., 2020; Canete et al., 2020) and multilingual models that were trained on a large corpus containing many different languages (Conneau et al., 2020; Xue et al., 2021). The training data of these multilingual models often largely consist of higher-resourced languages, but recent work has demonstrated that using *only* low-resourced languages can also achieve competitive performance (Ogueji et al., 2021).

Other work has also focused on analysing these large models, investigating multilingualism (K et al., 2020), syntactic transfer between languages (Dhar & Bisazza, 2018), or interpreting what information models use when predicting entity types (Agarwal et al., 2021).

Recently, Adelani et al. (2021) introduced a high-quality named entity recognition (NER) dataset for 10 low-resourced African languages and performed some analysis into which pre-trained models perform best, how domain transfer performs as well as the cross-lingual transfer capabilities of models. We analyse the latter aspect further, by investigating which features transfer between languages, how different training schemes affect performance and how robust the trained models are.

Our results show that when the source and target dataset contain many shared tokens, then transfer performance is generally higher. We also find that adaptively fine-tuning a multilingual model on unlabelled monolingual data can, depending on which languages are used, either improve transfer performance on downstream tasks or reduce it when the model overfits to a specific language. Finally, we demonstrate that when performing transfer learning, there is often a large variance in performance when varying the initial random seed – much larger than standard tasks where we fine-tune and

[1] https://anonymous.4open.science/r/NERTransferAnon-66D7

evaluate on the same language – possibly indicating that some aspects of transfer learning are not robust.

2 BACKGROUND

2.1 NAMED ENTITY RECOGNITION (NER)

Named Entity Recognition is a token classification task in which the objective is to classify each token as one of a few classes, e.g. person, location, date, organisation, or no entity. NER is an impactful field (Sang & Meulder, 2003; Lample et al., 2016) with many applications (Marrero et al., 2013), including information retrieval and spell-checking (Adelani et al., 2021).

2.2 TRANSFER LEARNING

Transfer learning is a technique that is often used in NLP to improve performance while requiring less task-specific data (Ruder et al., 2019). One common form of transfer starts by training a large language model on a massive corpus of unlabelled data, using these learned weights as the starting point for a specific problem, and fine-tuning further on task-specific labelled data (Ruder, 2021). The idea is that the pre-training process instils knowledge into the model about how language behaves on a general level, which then does not need to be learned from scratch using the smaller amount of task-specific data. Often, if the pre-training data is in a substantially different domain from the target task, we employ *adaptive fine-tuning*, which fine-tunes the pre-trained model on unlabelled data in the domain (Gururangan et al., 2020) or language (Pfeiffer et al., 2020) of the target task.

3 METHODOLOGY

This report builds upon the work of Adelani et al. (2021) and investigates the following questions:

1. How much does language-adaptive fine-tuning (using a language modelling objective) on different languages affect downstream performance after fine-tuning on task-specific data?
2. Which languages are the best for doing zero-shot transfer from?
3. What features or aspects get transferred between the languages we examine?

To answer the above questions, we fundamentally fine-tune different models on the MasakhaNER dataset (Adelani et al., 2021) and compare their performance. We consider all 10 languages, **Hau**sa, **Igbo** **Kin**yarwanda, **Lug**anda, **Luo**, Nigerian Pidgin - **pcm**, **Swa**hili, **Wol**of, **Yor**ùbá, **Amh**aric. For consistency, we refer to pre-training as any approach that trains a language model on a large, unlabelled corpus, whereas fine-tuning means taking a pre-trained model, and training it end-to-end on a smaller, labelled dataset. Additionally, we refer to fine-tuning a pre-trained model on an unlabelled corpus in another language using a language modelling objective as language-adaptive fine-tuning.

For point (1), we largely follow the approach of Adelani et al. (2021), and use this as a basis for the subsequent components. This involves fine-tuning different models on the language-specific NER data and analysing the effect of the language used for language-adaptive fine-tuning. For point (2) above, we consider how good zero-shot transfer is when fine-tuning on NER data from other languages by evaluating the models we trained for (1) on all 10 languages. To answer (3), we examine the statistical properties of the datasets as well as the contextual word embeddings obtained after various pre-training and fine-tuning steps.

4 EXPERIMENTS & RESULTS

4.1 EXPERIMENTAL SETUP

We fine-tune each model 5 times with different random seeds (to account for variability) and report the mean and standard deviation here. We use the MasakhaNER implementation[2] and use the same

[2] https://github.com/masakhane-io/masakhane-ner/

Table 1: Performance of different models after fine-tuning and evaluating on NER data. We use a Mann-Whitney U test (Mann & Whitney, 1947) as some data failed a Shapiro Wilks normality test (Shapiro & Wilk, 1965). * indicates a statistically significant difference ($p < 0.05$) between the base model and the one under consideration, **bold** implies * and being the maximum per language. The leftmost column shows the model we started with before fine-tuning on language-specific NER data, while the other columns indicate the NER fine-tuning and evaluation language. For example, base $\rightarrow X$ is the language adaptive model for each column.

Starting point for NER fine-tune	wol	pcm	yor	hau	ibo	luo	lug	kin	swa	amh
base $\rightarrow X$	66.9 (1.7)	87.1 (0.8)	**83.3 (0.3)***	**91.6 (0.4)***	**87.9 (0.5)***	76.2 (1.2)	**84.5 (0.5)***	**78.3 (1.0)***	**89.6 (0.6)***	**78.2 (0.8)***
base $\rightarrow$ swa	**67.3 (1.3)***	88.0 (0.8)	78.3 (1.0)	88.8 (0.2)*	84.3 (0.8)	77.2 (1.4)	82.0 (0.5)*	75.2 (1.0)	**89.6 (0.6)***	68.9 (0.9)
base	64.2 (1.3)	87.3 (0.9)	77.9 (0.3)	89.5 (0.4)	84.9 (0.7)	74.5 (1.3)	80.2 (0.7)	73.7 (0.7)	87.8 (0.5)	70.7 (1.1)

hyperparameters and language codes as Adelani et al. (2021). All metrics reported are overall F1 scores on the test set (to compare against prior work), using the 'begin' repair strategy as specified by Palen-Michel et al. (2021).

4.2 Language-adaptive fine-tuned Models

In this section, we determine the effect of using different language-adaptive fine-tuned models. Each of the models we consider is based on xlm-roberta (Conneau et al., 2020), as it demonstrated high performance and fast training (Adelani et al., 2021). The first model we use is called 'base', and it is simply xlm-roberta-base, downloaded from Huggingface[3]. The other models[4] we consider used xlm-roberta-base as their starting point, but additionally performed language-adaptive fine-tuning on a monolingual corpus and were shown to perform better on NER tasks (Adelani et al., 2021).

For each language X, we use 3 different models, base, base-Swahili and base-X, where the latter 2 were subject to further language-adaptive fine-tuning on their respective languages. The base-X models allow us to investigate how much the language-adaptive fine-tuning on the language in question improves downstream performance relative to the base model. The base-Swahili model, on the other hand, provides information on how downstream performance is affected by language-adaptive fine-tuning on a different, but potentially related (either geographically or lexically) language. We then fine-tune these models on NER data and report the results. For clarity and brevity, we sometimes contract the training procedure of a model, for example, base $\rightarrow$ hau $\rightarrow$ swa is the xlm-roberta-base model that performed language-adaptive fine-tuning on Hausa, and NER fine-tuning on Swahili.

The results are shown in Table 1, and the language adaptive fine-tuned models usually perform much better than the base model, with the Swahili model being in between. The standard deviations between the different seeds are quite large, however, so not all results are statistically significant (using a Mann-Whitney U test). In most cases, we replicate (within error bounds) Adelani et al. (2021) and Palen-Michel et al. (2021), with the single exception of Nigerian Pidgin that was fine-tuned from a language-adaptive model, possibly because of different model versions.

4.3 Cross-lingual Transfer

This experiment investigates fine-tuning one of the above models on one specific language (e.g. Yorùbá) and evaluating on another (e.g. Hausa). In particular, we take the models obtained from the previous section and evaluate them on all of the available languages.

These results are shown in Figure 1, with the y-axis representing the evaluation language, while the x-axis represents either the language we performed NER fine-tuning on (Figure 1a), the adaptive fine-tuning language (Figure 1c) or both (Figure 1b). Specifically, in Figure 1a, as expected, the diagonal is brighter than the off-diagonal elements, as fine-tuning on the same language one evaluates on improves scores significantly. In Figure 1b we see that performing language-adaptive fine-tuning improves the performance on that specific language. In transfer performance, however, this shows a mixed result, as for some language pairs, using a model that has been subject to language-adaptive fine-tuning on the same language as one fine-tunes on helps (e.g. the *pcm* column and *kin* row), but

[3] https://huggingface.co/xlm-roberta-base
[4] https://huggingface.co/Davlan

for others this effect is minor, or even negative (e.g. *yor* transferring to *lug*). We also notice that for other languages, notably Swahili and Hausa, using adaptively fine-tuned models (and then fine-tuning on NER data from the same language) diminishes the transfer capabilities from these languages (see the *hau* and *swa* columns in Figure 1g), possibly indicating overfitting. This is similar to what Pfeiffer et al. (2020) found when performing adaptive fine-tuning on the source language – transfer performance generally decreased.

In Figure 1c we use different adaptively fine-tuned models, but fine-tune on the same Swahili NER data, and again evaluate on each language. At first glance, horisontal lines can be seen, indicating that the adapted model does not affect the final score that much in this case, although the diagonal is usually slightly (e.g. *pcm, luo*) or significantly (e.g. *wol, lug, ibo, amh*) brighter. Again, we see a similar overfitting problem to the above when using a Swahili adapted model and fine-tuning on Swahili NER, resulting in much poorer transfer performance.

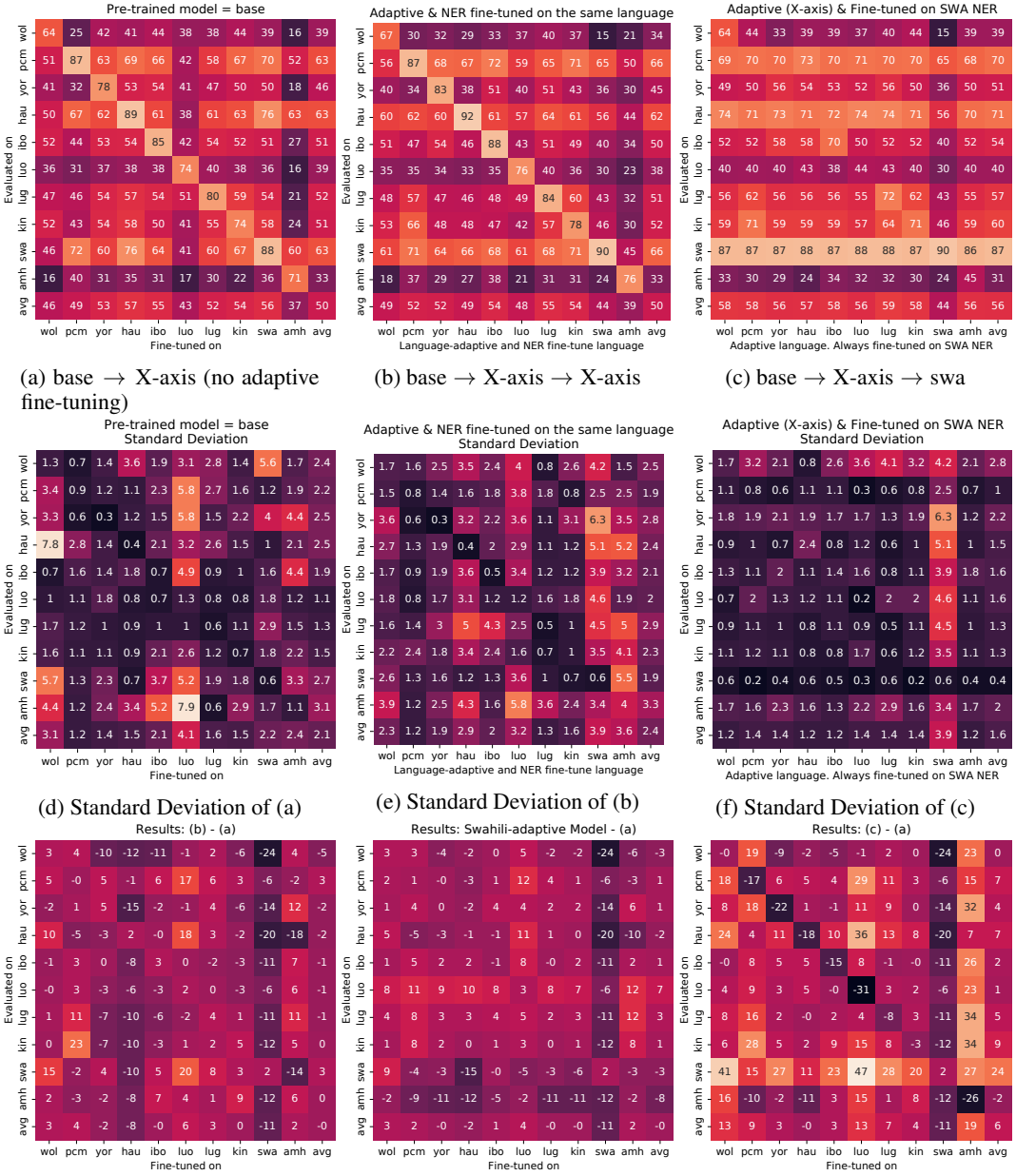
In Figures 1d, 1e and 1f we show the standard deviations over 5 seeds and the results seem to indicate that the standard deviation is generally higher when performing transfer as compared to performing standard evaluation on the fine-tuned language, which seems to indicate a lack of robustness to random initialisations.

Finally, in Figures 1g, 1h and 1i, we showcase the absolute performance gain of one method over another, to quickly be able to notice relative differences, either positive or negative.

We also consider the above in slightly more detail by looking at each NER category individually, to see if any perform much better or worse than the others. Figure 2 indicates that dates transfer quite poorly, particularly for Luo, while organisations transfer poorly for Amharic.

4.3.1 Data overlap

To attempt to explain some of the results shown in the previous section, here we examine the datasets a bit more carefully, specifically analysing the data overlap between different languages, and whether this has any correlation with the transfer performance. To do so, we investigate the overlap of each entity in the respective datasets. We call a token overlapping when the same token is labelled as the same entity type in two different datasets (e.g. John[NAME] would overlap with John[NAME], but would not overlap with John[ORG] Deere[ORG]). We count the number of overlaps from source language A (x-axis) to target language B (y-axis) as the number of occurrences of this token in A's dataset, as this could measure how much data the model can use from language A that might transfer to language B. We do not distinguish between tokens that are at the beginning of an entity or in the middle thereof (i.e. we consider B-PER and I-PER to be the same for this experiment). We consider the entire dataset, i.e. train + dev + test, to obtain a more representative sample, although this does not calculate the overlap between e.g. the train set of A and the test set of B. Other ways of calculating the overlap exist, like only considering unique entities (which we avoid as one entity overlapping multiple times is relevant) or considering the minimum occurrences of an overlapping token in either language. This was correlated with and similar to our approach, however, so it does not have a significant impact. Figure 3a shows the results, and a few things are immediately clear. Firstly, Wolof and Luo have much less data than the other languages, and thus much less overlap, potentially explaining why these two performed poorly in previous experiments. Secondly, there seems to be quite a lot of overlap in general. Swahili and Hausa also show many tokens in common, possibly due to Arabic influences on both of these (Versteegh, 2001), but it could also simply be e.g. international names and dates. We see a strong correlation (Pearson's coefficient = 0.73) between how many tokens overlap and the performance in Figure 3b. The procedure here was simply to compute the correlation between the data overlap (as in Figure 3a) and the performance when fine-tuning on one language and evaluating on another, starting from the pre-trained base model (as in Figure 1a). We do not consider the diagonal elements, as they contain the performance of evaluating on language X after fine-tuning on language X and are thus not considered as transfer learning. These results do not imply a causal relationship, however, as previous work (K et al., 2020) has shown that lexical overlap has a negligible impact on transfer performance, and word order and model depth contributes more. This might be specific to the task under consideration, however, as other work still (Lin et al., 2019) has shown that, for some tasks, the word and subword overlap between languages is a useful proxy for expected performance when performing cross-lingual transfer. Furthermore, we observe that Amharic, due to its different script, has zero overlap with any of the other languages, while still displaying non-trivial transfer, indicating that more intricate mechanisms are at play.



(a) base → X-axis (no adaptive fine-tuning)

(b) base → X-axis → X-axis

(c) base → X-axis → swa

(d) Standard Deviation of (a)

(e) Standard Deviation of (b)

(f) Standard Deviation of (c)

(g) Performance difference when adding language-adaptive fine-tuning. Swahili and Hausa transfer worse on average, while Luo improves.

(h) Performance difference between using a Swahili adaptively fine-tuned model and no language-adaptive fine-tuning after subsequent fine-tuning on NER data. Hausa fine-tuning performs much worse when evaluated on Swahili.

(i) Performance difference between language-adaptive fine-tuning on the X-axis plus fine-tuning on Swahili NER and just using the base model and fine-tuning on the X-axis NER. The Luo pre-trained model in particular has significant increase.

Figure 1: Heatmaps indicating the average performance over 5 seeds of specific models on specific languages (y-axis) after being fine-tuned on another language's NER data (x-axis). In general, we notice a large standard deviation, indicating that this process is unreliable. The bottom row shows the difference between one technique, and base, i.e. how much improvement does this new model give over using the base model. *avg* indicates the average per row or column respectively.

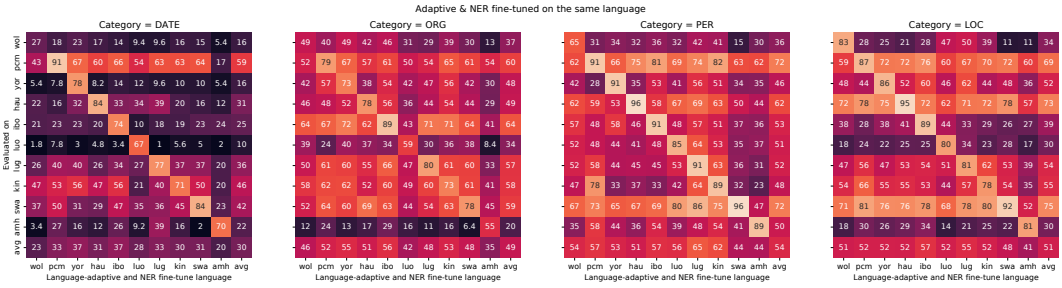
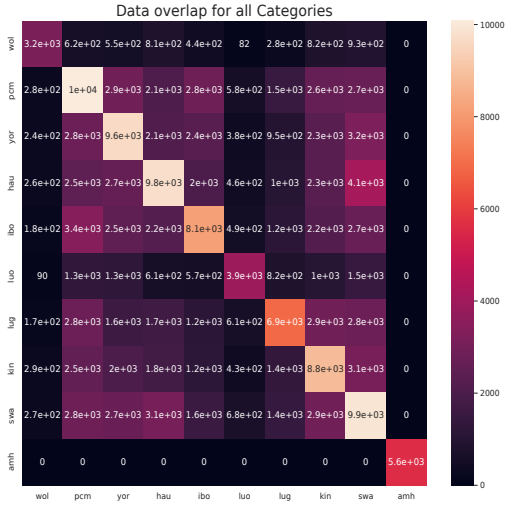
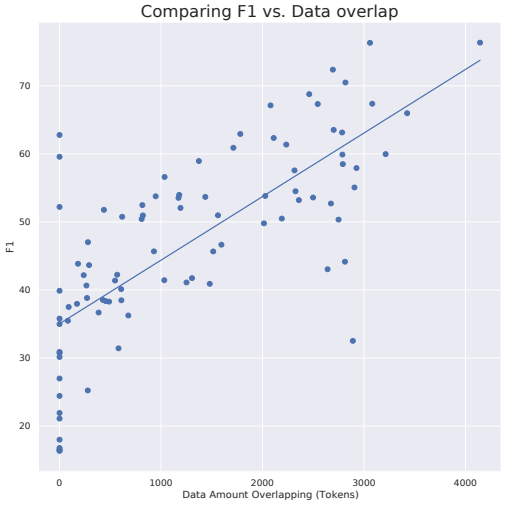


Figure 2: Heatmaps for the language-adaptive fine-tuned model (Figure 1b), broken down by category.



(a) Data Overlap. Row i, column j indicates the overlap if language j was the source (i.e. fine-tune language) and language i was the target (i.e. evaluation language).



(b) Correlation between data overlap and F1 performance when performing zero-shot transfer. 0.73 Pearson's correlation coefficient with $p < 0.05$.

Figure 3: Data overlap & effect on performance. We note that Amharic has no overlap with any of the other languages, as it uses a different script.

4.4 Representations

Our final experiment follows prior work by Hsu et al. (2019) by investigating the contextual word embeddings from the different models, specifically looking into how these embeddings change as we perform different fine-tuning operations. We take the last 4 layers from the language model (i.e. not the dense final layer) and use the sum of these hidden states to obtain a word vector (of size 768). We use the sentences from the dataset, and only extract the 4 different NER categories for computational reasons. We compute the mean vector per category, which we use in the following. To visualise the data, we show the results after performing PCA.

4.4.1 Variability

We found a large amount of variability when fine-tuning the models on different random seeds (see Figures 1d, 1e and 1f) and in this experiment we investigate the effect of using different initialisations on these embeddings.

Figure 4 shows the results for a few languages pairs, and immediately we can see that Figure 4a has clusters corresponding to the different categories, even when using different seeds. Figures 4b, 4c and 4d on the other hand cluster more toward seeds, so the categories differ when using different seeds. This could indicate that the Swahili model is more consistent and robust to random initialisations,

and learns roughly the same embeddings for each seed. On the other hand, when fine-tuning from Kinyarwanda, Luo or Wolof, there is no clear clustering of categories (despite a relatively large amount of data overlap between Kinyarwanda and Hausa), suggesting that these models cannot distinguish Hausa categories very well (possibly substantiated by the poorer results in Figure 1).

Now, the above analysis is somewhat impacted by the final linear layers in the models – it is entirely possible that two models that have different embeddings also have different final layers and end up classifying examples exactly the same. We can, however, still use these experiments to extract some qualitative information about the embeddings of different languages. Furthermore, Figures 4f and 4e – in which the language being investigated is the same as what the models trained on – contain results where the clustering is predominantly towards categories, bolstering the validity of this approach.

4.4.2 Different Languages

Here we consider the same model and analyse the differences in embeddings from different languages, and how this evolves. For example, in Figure 5a we see that for Nigerian Pidgin (which transferred quite well previously), the predominant clusters are again categories, and the same categories from different languages are grouped together.

4.4.3 Different Models

This experiment examines different models on the same language, specifically looking at what happens to these embeddings when a model is further fine-tuned. Figure 5b shows that performing fine-tuning on models does affect the embeddings quite significantly, although there does still seem to be a similar relative positioning between the categories - almost as if in PCA, one principal component was the model used, and another was the category.

5 Analysis, Discussion & Future Work

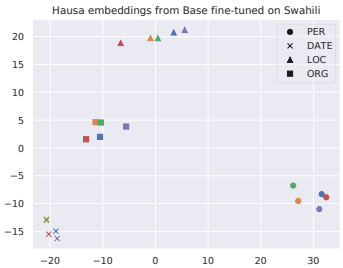
We touched on a few different topics in this report, all related to transfer learning and how this affects the F1 score. We found that language-adaptive fine-tuning has a positive effect on transfer performance, but that overfitting is a real risk, so the model that does best on one language often suffers in transfer to other languages, potentially motivating less overspecialised models in favour of more general models which would also hopefully be more robust. As usual, data is king, and having more data overlap is highly correlated with transfer performance, although we tested relatively high-quality datasets here, so this might not transfer well to low quality, noisy datasets (Alabi et al., 2020). Amharic also displayed non-trivial transfer, even though it shares no overlap with any of the other languages, so investigating this phenomenon further is a promising avenue. Future work could include looking at different languages, investigating transfer from a geographical (Adelani et al., 2021), language-family or linguistic distance perspective, or how combining two (or more) datasets might strike a balance between high single-task performance and generalisability.

6 Conclusion

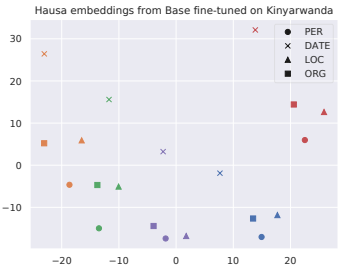
We considered different facets of transfer learning, namely how language-adaptive fine-tuning and fine-tuning on task-specific data in another language affects evaluation performance. We also investigated reasons for our observations, notably, data overlap between train and testing datasets, as well as the models' hidden states and embeddings. In some cases we found very large variances, making reliably performing transfer learning difficult. We answered our original 3 questions, specifically that (1) using a language-adaptive fine-tuned model improves performance on the corresponding language, possibly while reducing transfer performance. Regarding point (2), we found the best language to perform zero-shot transfer from does depend on various factors, although the amount of data overlap could help inform this choice, by choosing languages that have a large overlap. Finally, for (3), we found that overlapping tokens might be a large part of what is transferred, but some linguistic knowledge could also be transferred, between e.g. Swahili and Hausa, as clear separation of categories was apparent. In other cases (e.g. Luo and Hausa), a less clear clustering was observed.

Our main conclusions are that (in our case, with this dataset) using language-adaptive fine-tuned models usually improve performance over the base model after subsequent fine-tuning. We found

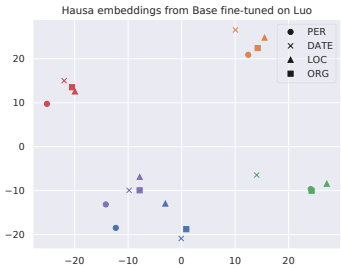
high levels of data overlap, and found a strong correlation between this and the F1 performance, although this does not imply causality.



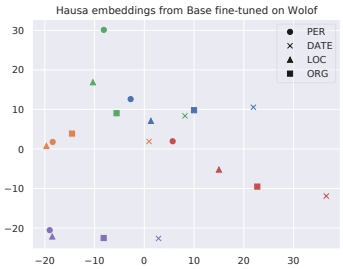
(a) Clear category clusters.



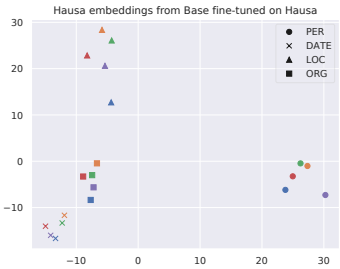
(b) Less clear clustering, but largely revolving around colours.



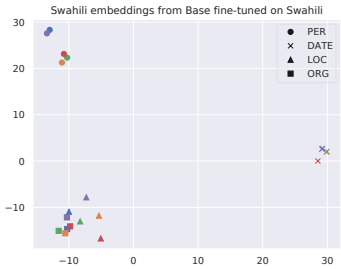
(c) Clear colour clusters.



(d) Clear colour clusters.

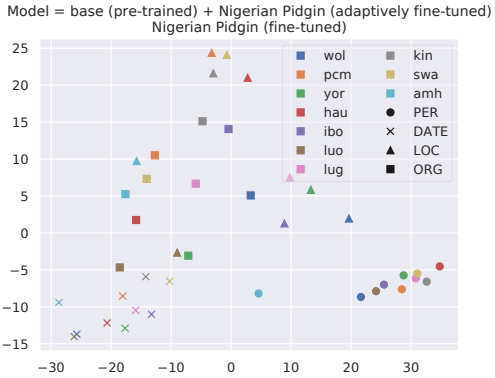


(e) Hausa embeddings from a model that was fine-tuned on Hausa – with clear category clusters.

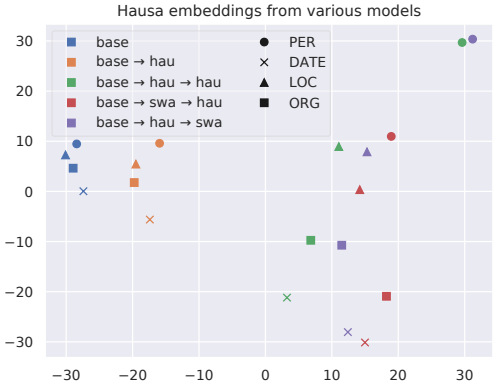


(f) Swahili embeddings from a model that was fine-tuned on Swahili – with clear category clusters, although locations and organisations are grouped close together.

Figure 4: Scatter plots of embeddings from different models, languages and categories. The shapes indicate different categories, whereas the colours indicate different starting points, i.e. seeds.



(a) Embeddings for different languages from a Nigerian Pidgin model, fine-tuned on Nigerian Pidgin. This clustering is quite robust to different initialisations and similar across different seeds.



(b) Hausa embeddings from different models. For example, base → hau → swa started with a base model, performed language-adaptive fine-tuning on unlabelled Hausa data and finally performed fine-tuning on Swahili NER data.

Figure 5: Embeddings of (a) multiple languages with one model and (b) Hausa embeddings from different models after performing PCA.

REFERENCES

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 O. 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. *Transactions of the Association for Computational Linguistics*, 9:1116–1131, 10 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00416. URL https://doi.org/10.1162/tacl_a_00416.

Oshin Agarwal, Yinfei Yang, Byron C. Wallace, and Ani Nenkova. Interpretability analysis for named entity recognition to understand system predictions and how they can improve. *Comput. Linguistics*, 47(1):117–140, 2021. doi: 10.1162/coli_a_00397. URL https://doi.org/10.1162/coli_a_00397.

Jesujoba Alabi, Kwabena Amponsah-Kaakyire, David Adelani, and Cristina España-Bonet. 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*, pp. 2754–2762, Marseille, France, May 2020. European Language Resources Association. URL https://www.aclweb.org/anthology/2020.lrec-1.335.

José Canete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. Spanish pre-trained bert model and evaluation data. *Pml4dc at iclr*, 2020:2020, 2020.

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 (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 8440–8451. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.747. URL https://doi.org/10.18653/v1/2020.acl-main.747.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. BERTje: A Dutch BERT Model. arXiv:1912.09582, December 2019. URL http://arxiv.org/abs/1912.09582.

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 (eds.), *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)*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1423. URL https://doi.org/10.18653/v1/n19-1423.

Prajit Dhar and Arianna Bisazza. Does syntactic knowledge in multilingual language models transfer across languages? In Tal Linzen, Grzegorz Chrupala, and Afra Alishahi (eds.), *Proceedings of the Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2018, Brussels, Belgium, November 1, 2018*, pp. 374–377. Association for Computational Linguistics, 2018. doi: 10.18653/v1/w18-5453. URL https://doi.org/10.18653/v1/w18-5453.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8342–8360, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.740. URL https://aclanthology.org/2020.acl-main.740.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. *CoRR*, abs/2103.03874, 2021. URL https://arxiv.org/abs/2103.03874.

Tsung-Yuan Hsu, Chi-Liang Liu, and Hung-yi Lee. Zero-shot reading comprehension by cross-lingual transfer learning with multi-lingual language representation model. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pp. 5932–5939. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1607. URL https://doi.org/10.18653/v1/D19-1607.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. SMART: robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 2177–2190. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.197. URL https://doi.org/10.18653/v1/2020.acl-main.197.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. Cross-lingual ability of multilingual BERT: an empirical study. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=HJeT3yrtDr.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In Kevin Knight, Ani Nenkova, and Owen Rambow (eds.), *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pp. 260–270. The Association for Computational Linguistics, 2016. doi: 10.18653/v1/n16-1030. URL https://doi.org/10.18653/v1/n16-1030.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. Choosing transfer languages for cross-lingual learning. In Anna Korhonen, David R. Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 3125–3135. Association for Computational Linguistics, 2019. doi: 10.18653/v1/p19-1301. URL https://doi.org/10.18653/v1/p19-1301.

H. B. Mann and D. R. Whitney. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50 – 60, 1947. doi: 10.1214/aoms/1177730491. URL https://doi.org/10.1214/aoms/1177730491.

Mónica Marrero, Julián Urbano, Sonia Sánchez-Cuadrado, Jorge Morato, and Juan Miguel Gómez-Berbís. Named entity recognition: fallacies, challenges and opportunities. *Computer Standards & Interfaces*, 35(5):482–489, 2013.

Mihai Masala, Stefan Ruseti, and Mihai Dascalu. Robert - A romanian BERT model. In Donia Scott, Núria Bel, and Chengqing Zong (eds.), *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pp. 6626–6637. International Committee on Computational Linguistics, 2020. doi: 10.18653/v1/2020.coling-main.581. URL https://doi.org/10.18653/v1/2020.coling-main.581.

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. URL https://aclanthology.org/2021.mrl-1.11.

Chester Palen-Michel, Nolan Holley, and Constantine Lignos. SeqScore: Addressing barriers to reproducible named entity recognition evaluation. In *Proceedings of the 2nd Workshop on*

Evaluation and Comparison of NLP Systems, pp. 40–50, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.eval4nlp-1.5.

Jonas Pfeiffer, Ivan Vulic, Iryna Gurevych, and Sebastian Ruder. MAD-X: an adapter-based framework for multi-task cross-lingual transfer. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 7654–7673. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.617. URL https://doi.org/10.18653/v1/2020.emnlp-main.617.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Sebastian Ruder. Recent Advances in Language Model Fine-tuning. http://ruder.io/recent-advances-lm-fine-tuning, 2021.

Sebastian Ruder, Matthew E Peters, Swabha Swayamdipta, and Thomas Wolf. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pp. 15–18, 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 (eds.), *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*, pp. 142–147. ACL, 2003. URL https://aclanthology.org/W03-0419/.

Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

Kees Versteegh. Linguistic contacts between arabic and other languages. *Arabica*, 48:470–508, 12 2001. doi: 10.1163/157005801323163825.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pp. 483–498. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.41. URL https://doi.org/10.18653/v1/2021.naacl-main.41.