*CLEAN*ANERCorp: Identifying and Correcting Incorrect Labels in the ANERcorp Dataset

Mashael Al-Duwais, Hend Al-Khalifa and Abdulmalik Al-Salman
King Saud University
Riyadh, Saudi Arabia
{malduwais, hendk, salman}@ksu.edu.sa

Abstract

Label errors are a common issue in machine learning datasets, particularly for tasks such as Named Entity Recognition. Such label errors might hurt model training, affect evaluation results, and lead to an inaccurate assessment of model performance. In this study, we dived deep into one of the widely adopted Arabic NER benchmark datasets (ANERcorp) and found a significant number of annotation errors, missing labels, and inconsistencies. Therefore, in this study, we conducted empirical research to understand these errors, correct them and propose a cleaner version of the dataset named CLEANANERCorp. CLEANANERCorp will serve the research community as a more accurate and consistent benchmark.

Keywords: Arabic NER, Label Error, Dataset.

1. Introduction

Named Entity Recognition (NER) is the task of identifying both spans and types of named entities in text. It is a fundamental task in the natural language processing pipeline.

The ANERcorp dataset is the most well-known and utilized dataset for Arabic NER (Benajiba et al., 2007), and is a crucial benchmark for evaluating Arabic NER approaches. ANERcorp consists of 316 manually annotated articles from the news domain.

Deep Learning approaches have achieved state-of-the-art performance in the ANERcorp dataset with F1-score (0.84, 0.88, 0.89, 0.91, 0.92) (Antoun et al., 2021a, 2021b; Khalifa & Shaalan, 2019; Al-Qurishi & Souissi, 2021; Alsaaran & Alrabiah, 2021) respectively.

While researchers have relied heavily on the ANERcorp as a benchmark dataset to evaluate Arabic NER models, none has considered the dataset quality. Label errors and inconsistency can have significant impact on evaluating machine learning algorithms. Detecting and correcting these errors is crucial for training accurate NER models, as the quality of the training data directly impacts the model's performance.

Moreover, previous experiments did not consider all tags during their experiments and used different data splits. This poses challenges in comparing NER approaches and analyzing their errors.

To address this issue, we present a thorough re-annotation effort that corrects **6.34%** of the label mistakes in the ANERcorp dataset and produces a cleaner version of the dataset named (CLEANANERCorp) that significantly improves annotation quality and consistency.

To the best of our knowledge, this is the first study that systematically handles label mistakes in the ANERcorp dataset. We conducted extensive experiments on both the original ANERcorp dataset and our corrected dataset CLEANANERCorp and achieved superior results.

The contributions of this study are as follows:

- We present CLEANANERCorp, a clean version of ANERcorp that includes corrected, consistent and reliable NER annotations in both splits, where (6.45%) of the training set and (6.16%) of the test set of the ANERcorp have been updated.
- We re-evaluated the popular Arabic NER models with CLEANANERCorp and achieved a marginally high increase with the F1 score results, which is about (7.23%).
- We re-evaluated the popular Cross-lingual NER models that achieved state-of-the-art performance with the corrected test set and achieved higher results.

CLEANANERCorp is publicly available to encourage the community to use it and to improve its quality further[1].

2. Related Work

The process of identifying incorrect labels in Named Entity Recognition (NER) dataset is common in the literature. These errors can occur due to human annotator mistakes or inconsistencies in the labeling guidelines. Previous studies have addressed label quality in NER datasets (Helgadóttir, Loftsson and Rögnvaldsson, 2014; Abudukelimu *et al.*, 2018; Stanislawek *et al.*, 2019; Wang *et al.*, 2019; Reiss *et al.*, 2020; Rücker and Akbik, 2023).

(Wang *et al.*, 2019) proposed a manually corrected test set of CoNLL2003 called (CoNLL++) were they identified label mistakes in about 5.38% test sentences. Likewise, (Reiss *et al.*, 2020) proposed a more error-free version of the CoNLL2003 dataset, were they identified errors in about 3.7% of the dataset. Recently, (Rücker and Akbik, 2023) proposed CLEANCONLL, where they corrected 7.0% of all labels in the English CoNLL2003 dataset using manual re-annotation and cross checking.

[1] Github link: https://github.com/iwan-rg/CLEANANERCorp

OSACT 2024 Workshop, pages 13–19

To the best of our knowledge, there is no previous attempt to investigate the quality and label errors in Arabic NER dataset (ANERcorp).

3. ANERcorp Overview

ANERcorp is one of the earliest and most widely adopted NER corpora for Arabic. It was published in 2007 and has since become the standard in the Arabic NER literature. ANERcorp comprises two corpora for training and one for testing. The total number of articles included 316 from different newspapers.

The dataset annotation guidelines followed in the ANERcorp dataset were based on MUC Conventions (Sang & De Meulder, 2003). Following this guideline, the dataset was tagged with four entities: *person (PER), location (LOC), organization (ORG), and miscellaneous (MISC)*. The tagging scheme is the inside–outside–beginning (IOB) scheme originally proposed by (Ramshaw and Marcus, 1999). Therefore, any word on the text should be annotated as one of the following tags:

- B-PER: The Beginning of the name of a person.[2]
- I-PER: The continuation (Inside) of the name of a person.
- B-LOC: The Beginning of the name of a location.
- I-LOC: The Inside of the name of a location.
- B-ORG: The Beginning of the name of an organization.
- I-ORG: The Inside of the name of an organization.
- B-MISC: The Beginning of the name of an entity that does not belong to any of the previous classes (miscellaneous).
- I-MISC: The Inside of the name of an entity that does not belong to any of the previous classes.
- O: The word is not a named entity (Other).

The dataset contains (150,286) tokens and (32,114) types, which makes the ratio of tokens to types is (4.67). The distributions of the different tags are listed in Table 1.

Class	Ratio
PER	39%
LOC	30.4%
ORG	20.6%
MISC	10%

Table 1 Ratio of phrases by classes

In 2020, the CAMeL Lab (Obeid *et al.*, 2020) released a new version of ANERcorp, where they split the data and performed minor corrections agreed upon with the original author.

[2] The original dataset used B-PERS instead of B-PER and I-PERS instead of I-PER in the annotation. We re-annotate the dataset with the same original tags in the dataset but refer to them as B-PER and I-PER in this paper.

The changes from the original dataset include the following:

- Correct minor tag spelling errors.
- Convert the middle periods (·) and bold periods (•) to regular periods (.).
- Remove the blank Unicode character (\u200F).
- Add sentence boundaries after sequences of one or more periods.
- Split the dataset sequentially. The sentences containing the first 5/6 of the words go to training, and the rest go to testing. The training split had 125,102 words, and the test split had 25,008 words.

However, no previous efforts have been made to correct tagging errors and mislabeling in the ANERcorp dataset. We have carefully reviewed the original ANERcorp and identified the different types of labeling errors. They are listed below with examples:

A. Label Inconsistency

Some tokens were tagged differently for each sentence. For example, (الدولارات، جنيه استرليني) has been tagged sometimes as MISC and sometimes as O. Also, (الضفة الغربية) has been tagged as LOC and O in different sentences.

B. Wrong Labels

In Figure 1, the word "المتحدة" has been tagged as B-ORG while it should be tagged as I-LOC.

Figure 1 An Example of a Wrong Label

C. MISC tag Ambiguity

As the dataset follows the same classes that were defined in the MUC-6 Conventions (Sang and De Meulder, 2003) (Organization, Location, Person, and Miscellaneous), the MISC tag was not covered correctly and many MISC entities were tagged as O.

D. Sentence Beginning Ambiguity

We noticed an ambiguity in the first words of many sentences where the correct label was not clear. Figure 2 shows an example of such a sentence where the word (برند) has been tagged as (B-PER) and the meaning of the word is not clear.

Figure 2 Sentence Beginning Ambiguity Example

E. Typographical Errors

In addition to tagging errors, we noticed some typographical errors in the dataset. The dataset was written in two columns, where each word was placed on a separate line with its tag. We encountered two words attached to each other in one line without space. For example: (فيهاالبلدان، التفسيرالنصى، ولمافشلت، وراءوالدهم، المصادرالتاريخية، إنسبعةعراقيين، أكبرمحافظة).

4. Reannotation Process

The reannotation process was conducted in four distinct phases.

4.1 Annotation Guideline Definition

The ANERcorp annotations are based on MUC-6 Conventions (Sang and De Meulder, 2003) guidelines. Following these guidelines, the dataset is tagged with four entities: Person (PER), Location (LOC), Organization (ORG), and Miscellaneous (MISC). As there are no clear documentation of the ANERcorp annotation guidelines, we have defined our own guidelines that follow MUC-6 published guidelines and suit Arabic language. For example, we consider prefixes to be part of the entity names. For example: (منظمة الامم المتحدة), (شركة النفط النيجيرية), (بورصة نيويورك).

We developed a special handling for ambiguities in the guidelines to resolve cases that were not clear during the revision. In most cases, we assigned a tag that matched the context of the sentence. Following (Rücker and Akbik, 2023), we decided to tag the national sport team with ORG instead of LOC (المنتخب المصري، المنتخب السعودي). Political houses were also tagged as LOC (البيت الأبيض، الكرملين).

We have noticed inconsistency in tagging the currency, sometimes as MISC and sometimes as O or LOC. Following CoNLL tagging, we decided to label the currency and physical units as O instead of MISC.

4.2 Automatic Error Detection with CLEANLAB

CLEANLAB[3] is a framework that automatically detects label issues in a machine learning dataset using confident learning (Wang and Mueller, 2022). This framework uses existing models to detect dataset problems that can be fixed to train even better models. We utilized CLEANLAB as a first round to check the number of issues in the dataset. We detected (1945) issues. These issues have been manually investigated and corrected.

4.3 Manual Re-annotation

An annotator was hired to manually re-annotate all the entities in the dataset. The annotator was provided with guidelines and encouraged to use search engines and Wikipedia for suspicious token spans. The dataset was split into nine files for ease of handling.

4.4 Final Revision

After the re-annotating process of all the tokens, a final round of revision has been conducted by the annotator and the author of the paper to resolve any ambiguity and inconsistency in the updated tags.

Finally, we corrected a total of **9518** label mistakes, which is approximately **6.34%** of the dataset.

5. Evaluation

5.1 Dataset Statistics

All labeling errors and typographical errors detected were resolved. The following subsections present some statistics on the data.

A. Label Distribution

Tables 2 and 3 compare the total count of annotated named entities and the distribution across the four classes for CLEANANERCorp and the original ANERcorp. We observe that CLEANANERCorp has a slightly higher number of ORG and MISC entities than the base version. This originates from a more consistent use of ORG labels in organization names and MISC labels for adjectives and entity types, such as sports leagues and events.

	ANERcorp		**CLEANANERCorp**	
Class	#	%	#	%
PER	1499	5.99%	1504	6.01%
LOC	751	3.00%	813	3.25%
ORG	725	2.90%	1006	4.02%
MISC	400	1.60%	1081	4.32%
O	21633	86.50%	20604	82.39%
Total	25008	100%	25008	100%

Table 2 Statistics of test set entities in ANERcorp vs. CLEANANERCorp datasets.

	ANERcorp		**CLEANANERCorp**	
Class	#	%	#	%
PER	4926	3.94%	4906	3.92%
LOC	4301	3.44%	4649	3.72%
ORG	2691	2.15%	4254	3.40%
MISC	1263	1.01%	5278	4.22%
O	111921	89.46%	106015	84.74%
Total	125102	100%	125102	100%

Table 3 Statistics of entities of the training set in ANERcorp vs. CLEANANERCorp datasets.

B. Labels Changed

Table 4 shows the extent of the label updates introduced compared to the original dataset. A total of (**9518**) labels were modified from the original dataset, which is (**6.34%**) of the total dataset. Tables 5 and 6 further examine the update details for each data split.

[3] https://github.com/cleanlab

	TRAIN		TEST		TOTAL	
	#	%	#	%	#	%
Chan-ged	7974	**6.37**	1544	**6.17**	9518	**6.34**
Unch-anged	117128	93.6	23464	93.83	140592	93.66
Total	125102	100	25008	100	150110	100

Table 4 NER labels updated in CLEANANERCorp datasets.

	CLEANANERCorp Train Set	
	#	%
Label Corrected	1664	1.33%
Label Added	6310	5.04%
Label Unchanged	117128	93.63%
#Entities	125102	100%

Table 5 NER labels in the CLEANANERCorp train set according to the type of change.

	CLEANANERCorp Test Set	
	#	%
Label Corrected	392	1.57%
Label Added	1152	4.61%
Label Unchanged	23464	93.83%
#Entities	25008	100%

Table 6 NER labels in the CLEANANERCorp test set according to the type of change.

6. Experiments

To determine the extent to which our relabeling effort affects model performance, we re-evaluated a set of NER models on CLEANANERCorp and ANERcorp in two different settings: monolingual and cross-lingual transfer.

Currently, fine-tuning large pre-trained language models has achieved state-of-the-art performance on both monolinguals (Antoun, Baly and Hajj, 2021a, 2021b) and cross-lingual NER (Hu *et al.*, 2020; Lan *et al.*, 2020). Therefore, we selected pre-trained language models from the literature that report state-of-the-art results on Arabic and English-Arabic cross-lingual transfer and re-evaluated them on different dataset versions for the NER task.

For the cross-lingual transfer, we experimented with a zero-shot cross-lingual transfer from English to Arabic, where the model was trained on English data and tested on Arabic. We used the CoNLL2003 dataset (Sang and De Meulder, 2003) for training and validation.

Although there are other published results (Abdul-Mageed et al., 2021; Khalifa & Shaalan, 2019) with higher SOTA, they reported the results on different data splits and tested the models without the MISC tag, focusing only on three tags: person (PER), location (LOC), and organization (ORG), while setting other labels to the unnamed entity (O).

6.1 Reference Models

We re-evaluated state-of-the-art Arabic and multilingual language models on the CLEANANERCorp and ANERcorp datasets.

For the Arabic pretrained language models, we re-evaluated the following:

- **ARABERT**v0.2 base (Antoun, Baly and Hajj, 2021a): The state-of-the-art Arabic-specific BERT model for various Arabic IE tasks. The model contained 24 layers of encoders stacked on top of each other, 16 self-attention heads, and a hidden size of 1024.
- **ARBERT** (Abdul-Mageed, Elmadany and Nagoudi, 2021): Arabic-specific Transformer LMs pre-trained on very large and diverse datasets, including MSA as well as Arabic dialects.
- **AraELECTRA** (Antoun, Baly and Hajj, 2021b): A pretrained ELECTRA model on a large-scale Arabic dataset.

For the cross-lingual experiments, we re-evaluated

- **mBERT** (Devlin *et al.*, 2019): Multilingual BERT pretrained on Wikipedia of 104 languages using masked language modelling (MLM).
- **XLM-RoBERT** (XLM-R) (Conneau *et al.*, 2020): A transformer-based multilingual masked language model pre-trained on text in 100 languages that obtains state-of-the-art performance on different cross-lingual tasks.
- **GigaBERT** (Lan *et al.*, 2020): A bilingual BERT for English-to-Arabic cross-lingual transfer trained on newswire English and Arabic text from the Gigaword dataset in addition to Wikipedia and Web crawl data.

Hyperparameter: For monolingual fine-tuning experiments, we followed the same hyperparameter reported by (Antoun et al., 2021b), where all the models were fine-tuned with batch size set to (32), maximum sequence length of (256), and learning rates (5e-5). For cross-lingual fine-tuning, we followed the same hyperparameters reported by (Hu et al., 2020), where mBERT was fine-tuned for two epochs, with a training batch size of (32) and a learning rate of (2e-5), and XLM-R was fine-tuned for two epochs with a learning rate of 3e-5 and size of 16. All hyperparameter tuning for the cross-lingual experiment was performed on the English validation data.

6.2 Monolingual Results

The experimental results of the tested models for the different dataset versions are listed in Table 7. F1-score was averaged over three runs with different seeds for each experimental setting.

Model	Train/Test : **ANERcorp**	Train/Test : **CLEANANERCorp**
AraBERT v2	0.83	**0.89**
ARBERT	0.83	**0.89**
AraELECTRA	0.82	**0.87**

Table 7 Average F1 score of fine-tuning Arabic LMs on ANERcorp vs. CLEANANERCorp datasets.

The results show that CLEANANERCorp achieved marginally higher performance on all tested models compared to the original dataset, which indicates that our relabeling effort successfully improved label quality and consistency.

AraBERT F1 score has increased by (7.23%) from (0.83) to (0.89) after re-annotation. Table 8 shows a detailed comparison of each entity type in terms of Precision, Recall and F1-score for the AraBERT model on the two versions of the datasets.

We can see that all the F1 scores increased after correction, and the highest gain in entity F1 score was from the MISC and ORG labels, where the F1 score increased by (26.47%) and (16%), respectively.

	ANERcorp			**CLEANANERCorp**		
	Prec	**Rec**	**F1**	**Prec**	**Rec**	**F1**
LOC	0.89	0.93	0.91	0.94	0.92	0.93
MISC	0.73	0.63	0.68	0.85	0.86	0.86
ORG	0.76	0.73	0.75	0.85	0.87	0.86
PER	0.88	0.84	0.86	0.93	0.90	0.92
Overall	0.84	0.82	0.83	0.89	0.89	0.89

Table 8 Entity-based precision, recall, and F1 score of fine-tuned AraBERT on ANERcorp vs. CLEANANERCorp datasets.

6.3 Cross-Lingual Zero-Shot Transfer Results

Table 9 reports the average F1 scores over three runs with different seeds for each experimental setting.

From the results in Table 9, we can observe a high increase in F1 scores when transferring to the corrected dataset compared to those on the original test set.

Model	**Train: Conll2003 Test: ANERcorp**	**Train: Conll2003 Test: CLEANANERCorp**
mBERT-base	0.46	**0.48**
XLM-R-base	0.52	**0.62**
XLM-R-Large	0.53	**0.62**
GigaBERT	0.61	**0.72**

Table 9 Average F1 Scroe of Cross-lingual transfer on the ANERcorp vs. CLEANANERCorp datasets.

For example, fine-tuning XLM-R-base achieved (19.23%) increase from the (0.52) to (0.62) F1-score. Table 10 shows the F1 score per entity type, where we can see a high increase in the MISC label F1 score from (0.08) to (0.57), which justifies the increase in the overall score.

	ANERcorp			**CLEANANERCorp**		
	Prec	**Rec**	**F1**	**Prec**	**Rec**	**F1**
LOC	0.63	0.72	0.68	0.61	0.70	0.65
MISC	0.05	0.19	0.08	0.59	0.56	0.57
ORG	0.41	0.54	0.46	0.44	0.53	0.48
PERS	0.61	0.71	0.66	0.70	0.70	0.70
Overall	0.43	0.63	0.51	0.59	0.63	0.61

Table 10 Entity-based precision, recall, and F1 score of the Cross-lingual Transfer of XLM-R on ANERcorp vs. CLEANANERCorp dataset.

The above results indicate that CLEANANERCorp is more consistent with the CONLL2003 dataset and can be used to reflect the accuracy of the Cross-lingual Zero-Shot models more stably.

To get further insight into the label quality of the corrected and original dataset. We analyze the best model performance on cross-lingual zero-shot experiment using GigaBERT model.

Figures 3 and 4 show the classification report for the cross-lingual transfer on GigaBERT using the original and corrected version dataset. We noticed an improvement in the F1 score of all the tags specially for the MISC, PER and ORG were they have been mainly corrected in the new dataset version.

	precision	recall	f1-score	support
LOC	0.72	0.84	0.78	676
MISC	0.06	0.21	0.10	243
ORG	0.58	0.63	0.60	459
PER	0.75	0.76	0.75	906
micro avg	0.53	0.70	0.60	2284
macro avg	0.53	0.61	0.56	2284
weighted avg	0.64	0.70	0.66	2284

Figure 3 Classification Report for the Cross-lingual Zero Shot transfer using the original ANERcorp

	precision	recall	f1-score	support
LOC	0.73	0.84	0.78	682
MISC	0.70	0.68	0.69	835
ORG	0.69	0.57	0.63	590
PER	0.79	0.81	0.80	902
micro avg	0.74	0.73	0.73	3009
macro avg	0.73	0.72	0.72	3009
weighted avg	0.73	0.73	0.73	3009

Figure 4 Classification Report for the Cross-lingual Zero Shot transfer using CLEANANERCorp

Figure 5 shows the confusion matrix of the original dataset, where we see that the beginning of a person and location is often confused with the inside of MISC token. Figure 6 shows the confusion matrix of the corrected dataset with major improvements.

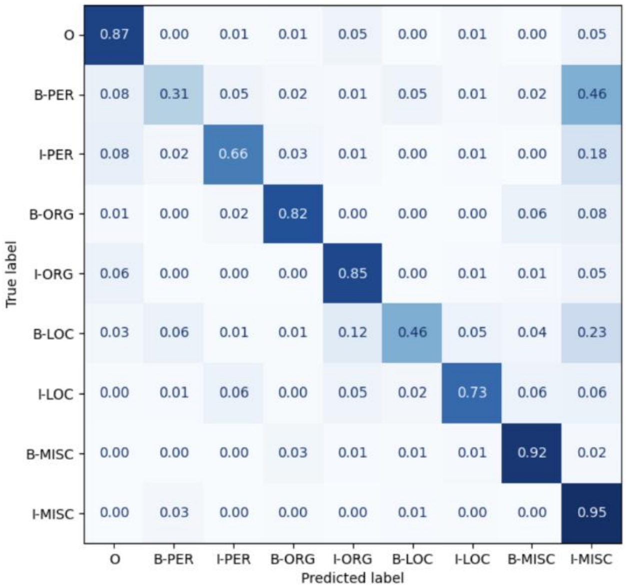

Figure 5 Confusion Matrix for Zero-shot Cross-lingual transfer using ANERCrop

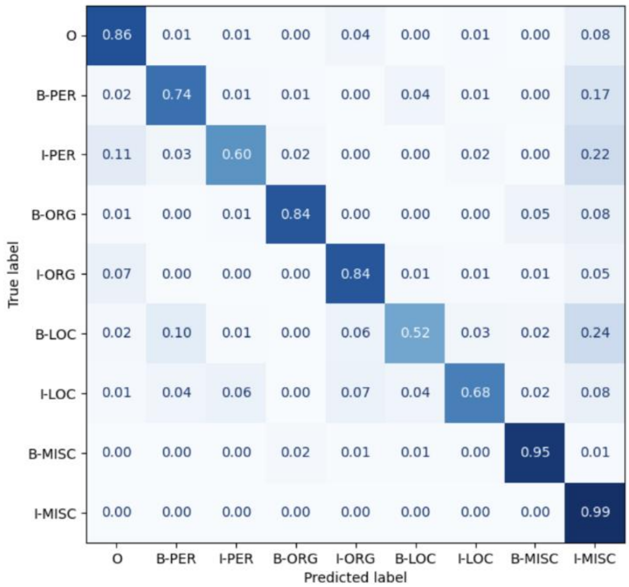

Figure 6 Confusion Matrix for Zero-shot Cross-lingual transfer using CLEANANERCorp

7. Conclusion

We presented CLEANANERCorp, a corrected and cleaner version of the widely adopted Arabic NER benchmark dataset ANERcorp. Our re-annotation updated (6.34%) the labels in the original dataset. Our evaluation of monolingual and cross-lingual NER language models achieved higher performance and strongly indicated that the overall annotation quality and consistency were significantly improved. Therefore, we contribute to improving the quality of the public Arabic NER datasets with updated and more consistent NER labels.

8. References

Abdul-Mageed, M., Elmadany, A. and Nagoudi, E.M.B. (2021) 'ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic', in C. Zong et al. (eds) *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. *ACL-IJCNLP 2021*, Online: Association for Computational Linguistics, pp. 7088–7105. Available at: https://doi.org/10.18653/v1/2021.acl-long.551.

Abudukelimu, H. *et al.* (2018) 'Error Analysis of Uyghur Name Tagging: Language-specific Techniques and Remaining Challenges', in N. Calzolari et al. (eds) *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. *LREC 2018*, Miyazaki, Japan: European Language Resources Association (ELRA). Available at: https://aclanthology.org/L18-1700 (Accessed: 30 March 2024).

Al-Qurishi, M.S. and Souissi, R. (2021) 'Arabic Named Entity Recognition Using Transformer-based-CRF Model', in M. Abbas and A.A. Freihat (eds) *Proceedings of the 4th International Conference on Natural Language and Speech Processing (ICNLSP 2021)*. *ICNLSP 2021*, Trento, Italy: Association for Computational Linguistics, pp. 262–271. Available at: https://aclanthology.org/2021.icnlsp-1.31 (Accessed: 25 February 2024).

Alsaaran, N. and Alrabiah, M. (2021) 'Arabic named entity recognition: A BERT-BGRU approach', *Comput. Mater. Contin*, 68, pp. 471–485.

Antoun, W., Baly, F. and Hajj, H. (2021a) 'AraBERT: Transformer-based Model for Arabic Language Understanding'. arXiv. Available at: https://doi.org/10.48550/arXiv.2003.00104.

Antoun, W., Baly, F. and Hajj, H. (2021b) 'AraELECTRA: Pre-Training Text Discriminators for Arabic Language Understanding'. arXiv. Available at: http://arxiv.org/abs/2012.15516 (Accessed: 25 February 2024).

Benajiba, Y., Rosso, P. and BenedíRuiz, J.M. (2007) 'ANERsys: An Arabic Named Entity Recognition System Based on Maximum Entropy', in A. Gelbukh (ed.) *Computational Linguistics and Intelligent Text*

Processing. Berlin, Heidelberg: Springer (Lecture Notes in Computer Science), pp. 143–153. Available at: https://doi.org/10.1007/978-3-540-70939-8_13.

Conneau, A. *et al.* (2020) ‘Unsupervised Cross-lingual Representation Learning at Scale’. arXiv. Available at: https://doi.org/10.48550/arXiv.1911.02116.

Devlin, J. *et al.* (2019) ‘BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding’. arXiv. Available at: https://doi.org/10.48550/arXiv.1810.04805.

Helgadóttir, S., Loftsson, H. and Rögnvaldsson, E. (2014) ‘Correcting Errors in a New Gold Standard for Tagging Icelandic Text’, in N. Calzolari et al. (eds) *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*. *LREC 2014*, Reykjavik, Iceland: European Language Resources Association (ELRA), pp. 2944–2948. Available at: http://www.lrec-conf.org/proceedings/lrec2014/pdf/677_Paper.pdf (Accessed: 30 March 2024).

Hu, J. *et al.* (2020) ‘XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalisation’, in *Proceedings of the 37th International Conference on Machine Learning*. *International Conference on Machine Learning*, PMLR, pp. 4411–4421. Available at: https://proceedings.mlr.press/v119/hu20b.html (Accessed: 20 December 2023).

Khalifa, M. and Shaalan, K. (2019) ‘Character convolutions for Arabic Named Entity Recognition with Long Short-Term Memory Networks’, *Computer Speech & Language*, 58, pp. 335–346. Available at: https://doi.org/10.1016/j.csl.2019.05.003.

Lan, W. *et al.* (2020) ‘An Empirical Study of Pre-trained Transformers for Arabic Information Extraction’, in B. Webber et al. (eds) *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. *EMNLP 2020*, Online: Association for Computational Linguistics, pp. 4727–4734. Available at: https://doi.org/10.18653/v1/2020.emnlp-main.382.

Obeid, O. *et al.* (2020) ‘CAMeL Tools: An Open Source Python Toolkit for Arabic Natural Language Processing’, in N. Calzolari et al. (eds) *Proceedings of the Twelfth Language Resources and Evaluation Conference*. *LREC 2020*, Marseille, France: European Language Resources Association, pp. 7022–7032. Available at: https://aclanthology.org/2020.lrec-1.868 (Accessed: 19 February 2024).

Ramshaw, L.A. and Marcus, M.P. (1999) ‘Text Chunking Using Transformation-Based Learning’, in S. Armstrong et al. (eds) *Natural Language Processing Using Very Large Corpora*. Dordrecht: Springer Netherlands (Text, Speech and Language Technology), pp. 157–176. Available at: https://doi.org/10.1007/978-94-017-2390-9_10.

Reiss, F. *et al.* (2020) ‘Identifying Incorrect Labels in the CoNLL-2003 Corpus’, in R. Fernández and T. Linzen (eds) *Proceedings of the 24th Conference on Computational Natural Language Learning*. *CoNLL 2020*, Online: Association for Computational Linguistics, pp. 215–226. Available at: https://doi.org/10.18653/v1/2020.conll-1.16.

Rücker, S. and Akbik, A. (2023) ‘CleanCoNLL: A Nearly Noise-Free Named Entity Recognition Dataset’. arXiv. Available at: http://arxiv.org/abs/2310.16225 (Accessed: 28 November 2023).

Sang, E.F.T.K. and De Meulder, F. (2003) ‘Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition’. arXiv. Available at: http://arxiv.org/abs/cs/0306050 (Accessed: 20 February 2024).

Stanislawek, T. *et al.* (2019) ‘Named Entity Recognition - Is There a Glass Ceiling?’, in M. Bansal and A. Villavicencio (eds) *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. *CoNLL 2019*, Hong Kong, China: Association for Computational Linguistics, pp. 624–633. Available at: https://doi.org/10.18653/v1/K19-1058.

Wang, W.-C. and Mueller, J. (2022) ‘Detecting Label Errors in Token Classification Data’. arXiv. Available at: http://arxiv.org/abs/2210.03920 (Accessed: 26 February 2024).

Wang, Z. *et al.* (2019) ‘CrossWeigh: Training Named Entity Tagger from Imperfect Annotations’, in K. Inui et al. (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)*. *EMNLP-IJCNLP 2019*, Hong Kong, China: Association for Computational Linguistics, pp. 5154–5163. Available at: https://doi.org/10.18653/v1/D19-1519.