Arabic Fine-Grained Entity Recognition

Haneen Abdallatif Liqreina
Birzeit University
Birzeit, Palestine
1195325@student.birzeit.edu

Mustafa Jarrar
Birzeit University
Birzeit, Palestine
mjarrar@birzeit.edu

Mohammed Khalilia
Birzeit University
Birzeit, Palestine
mkhalilia@birzeit.edu

Ahmed Oumar El-Shangiti
MBZUAI
Abu Dhabi, United Arab Emirates
ahmed.oumar@mbzuai.ac.ae

Muhammad Abdul-Mageed
UBC and MBZUAI
Vancouver, Canada
muhammad.mageed@ubc.ca

Abstract

Traditional NER systems are typically trained to recognize coarse-grained entities, and less attention is given to classifying entities into a hierarchy of fine-grained lower-level subtypes. This article aims to advance Arabic NER with fine-grained entities. We chose to extend Wojood (an open-source Nested Arabic Named Entity Corpus) with subtypes. In particular, four main entity types in Wojood, geopolitical entity (GPE), location (LOC), organization (ORG), and facility (FAC), are extended with 31 subtypes. To do this, we first revised Wojood's annotations of GPE, LOC, ORG, and FAC to be compatible with the LDC's ACE guidelines, which yielded $5,614$ changes. Second, all mentions of GPE, LOC, ORG, and FAC ($\sim 44K$) in Wojood are manually annotated with the LDC's ACE subtypes. We refer to this extended version of Wojood as $Wojood_{Fine}$. To evaluate our annotations, we measured the inter-annotator agreement (IAA) using both Cohen's Kappa and F_1 score, resulting in 0.9861 and 0.9889, respectively. To compute the baselines of $Wojood_{Fine}$, we fine-tune three pre-trained Arabic BERT encoders in three settings: flat NER, nested NER and nested NER with subtypes and achieved F_1 score of 0.920, 0.866, and 0.885, respectively. Our corpus and models are open-source and available at https://sina.birzeit.edu/wojood/.

1 Introduction

Named Entity Recognition (NER) is the task of identifying and classifying named entities in unstructured text into predefined categories such as people, organizations, locations, disease names, drug mentions, among others (li et al., 2020). NER is widely used in various applications such as information extraction and retrieval (Jiang et al., 2016), question answering (Liu et al., 2020), word sense disambiguation (Jarrar et al., 2023a; Al-Hajj and Jarrar, 2021), machine translation (Jain et al., 2019; Khurana et al., 2022), automatic summarization (Summerscales et al., 2011; Khurana et al., 2022), interoperability (Jarrar et al., 2011) and cybersecurity (Tikhomirov et al., 2020).

Traditional NER systems are typically trained to recognize coarse and high-level categories of entities, such as person (PERS), location (LOC), geopolitical entity (GPE), or organization (ORG). However, less attention is given to classifying entities into a hierarchy of fine-grained lower-level subtypes (Zhu et al., 2020; Desmet and Hoste, 2013). For example, locations (LOC) like Asia and Red Sea could be further classified into Continent and Water-Body, respectively. Similarly, organizations like Amazon, Cairo University, and Sphinx Cure can be classified into commercial, educational, and health entities, respectively. Belgium, Beirut, and Brooklyn can be classified into Country, Town, and Neighborhood instead of classifying them all as GPE. The importance of classifying named entities into subtypes is increasing in many application areas, especially in question answering, relation extraction, and ontology learning (Lee et al., 2006).

As will be discussed in the following sub-section, the number of NER datasets that support subtypes is limited, particularly for the Arabic language. The only available Arabic NER corpus with subtypes is the LDC's ACE2005 (Walker et al., 2005). However, this corpus is expensive. In addition, ACE2005 was collected two decades ago and hence may not be representative of the current state of Arabic language use. This is especially the case since language models are known to be sensitive to temporal and domain shifts (see section 5).

To avoid starting from scratch, we chose to extend upon a previously published and open-source Arabic NER corpus known as 'Wojood' (Jarrar et al., 2022). Wojood consists of $550K$ tokens manually annotated with 21 entity types. In particular, we manually classify four main entity types in Wojood (GPE, LOC, ORG, and FAC) with 31 new fine-grained subtypes. This extension is not straightforward as we have to change ($5,614$ changes) the original annotation of these four types of entities to align with LDC guidelines before extending them with subtypes. The total number of tokens that are annotated with the 31 subtypes is 47.6K.

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 310–323

Our extended version of Wojood is hereafter called *Wojood*$_{Fine}$. We measure inter-annotator agreement (IAA) using both Cohen's Kappa and F_1, resulting in 0.9861 and 0.9889, respectively.

To compute the baselines for *Wojood*$_{Fine}$, we fine-tune three pre-trained Arabic BERT encoders across three settings: (i) flat, (ii) nested without subtypes, and (iii) nested with subtypes, using multi-task learning. Our models achieve 0.920, 0.866, and 0.885 in F_1, respectively.

The remaining of the paper is organized as follows: Section 2 overviews related work, and Section 3 presents the *Wojood*$_{Fine}$ corpus, the annotation process, and the inter-annotator-agreement measures. In Section 4, we present the experiments and the fine-tuned NER models. In Section 5 we present error analysis and out-of-domain performance and we conclude in Section 6.

2 Related Work

Most of the NER research is focused on coarse-grained named entities and typically targets a limited number of categories. For example, Chinchor and Robinson (1997) proposed three classes: person, location and organization. The Miscellaneous class was added to CoNLL-2003 (Sang and De Meulder, 2003). Additional four classes (geo-political entities, weapons, vehicles, and facilities) were also introduced in the ACE project (Walker et al., 2005). The OntoNotes corpus is more expressive as it covers 18 types of entities (Weischedel et al., 2013).

Coarse-grained NER is a good starting point for named entity recognition, but it is not sufficient for tasks that require a more detailed understanding of named entities (Ling and Weld, 2012; Hamdi et al., 2021).

Substantial research has been undertaken to identify historical entities. For instance, the HIPE shared task (Ehrmann et al., 2020a) focused on extracting named entities from historical newspapers written in French, German, and English. One of its subtasks was the recognition and classification of mentions according to finer-grained entity types. The corpus used in the shared task consists of tokens annotated with five main entity types and 12 subtypes, following the IMPRESSO guidelines (Ehrmann et al., 2020b). A similar corpus, called NewsEye, was collected from historical newspapers in four languages: French, German, Finnish, and Swedish (Hamdi et al., 2021). The corpus is annotated with four main types: PER, LOC, ORG, and PROD. The LOC entities were further classified into five subtypes, and the ORG entities into two subtypes. Desmet and Hoste (2013) proposed a one million fine-grained NER corpus for Dutch, which was annotated using six main entity types and 27 subtypes (10 subtypes for PERS, three for ORG, nine for LOC, three for PROD, and two for events).

Zhu et al. (2020) noted that NER models cannot effectively process fine-grained labels with more than 100 types. Thus, instead of having many fine-grained entities at the top level, they propose a tagging strategy in which they use 15 main entity types and 131 subtypes. Additionally, Ling and Weld (2012) proposed a fine-grained set of 112 tags and formulated the tagging problem as multi-class multi-label classification.

A recent shared task was organized by Fetahu et al. (2023) at SemEval-2023 Task 2, called Multi-CoNER 2 (Fine-grained Multilingual Named Entity Recognition). A multilingual corpus (MULTICONER V2) was extracted from localized versions of Wikipedia covering 12 languages - Arabic is not included. The corpus was annotated with a NER taxonomy consisting of 6 coarse-grained types and 33 fine-grained subtypes (seven subtypes for Person, seven for Group, five for PROD, five for Creative Work, and five for Medical). Most participating systems outperformed the baselines by about 35% F_1.

There are a few Arabic NER corpora (Darwish et al., 2021), but all of them are coarse-grained. The ANERCorp corpus covers four entity types (Benajiba et al., 2007), CANERCorpus covers 14 religion-specific types (Salah and Zakaria, 2018), and Ontonotes covers 18 entities (Weischedel et al., 2013). The multilingual ACE2005 corpus (Walker et al., 2005), which includes Arabic, covers five coarse-grained entities and 35 fine-grained subtypes (3 subtypes for PERS, 11 for GPE, seven for LOC, nine for ORG, and five for FAC). Nevertheless, the ACE2005 corpus is costly and covers only one domain (media articles) that was collected 20 years ago. The most recent Arabic NER corpus is Wojood (Jarrar et al., 2022), which covers 21 nested entity types covering multiple domains. However, Wojood is a coarse-grained corpus and does not support entity subtypes.

To build on previous research on Arabic NER, we chose to extend the Wojood corpus with finer-grained subtypes. To ensure that our Wojood exten-

sion is compatible with other corpora, we chose to follow the ACE annotation guidelines.

3 *Wojood*$_{Fine}$ Corpus

Wojood$_{Fine}$ expands the annotation of the Wojood corpus (Jarrar et al., 2022), by adding fine-grain annotations for named-entity subtypes. Wojood is a NER corpus with 550K tokens annotated manually using 21 entity types. About 80% of Wojood was collected from MSA articles, while the 12% was collected from social media in Palestinian and Lebanese dialects (Curras and Baladi corpora (Haff et al., 2022; Jarrar et al., 2017, 2014)). One novelty of Wojood is its nested named entities, but some entity types can be ambiguous, which will affect downstream tasks such as information retrieval. For instance, the entity type "Organization" may refer to the government, educational institution, or a hospital to name a few. That is why *Wojood*$_{Fine}$ adds subtypes to four entity types: Geopolitical Entity (GPE), Organization (ORG), Location (LOC), and Facility (FAC). Table 3.3 shows the overall counts of the main four entity types in Wojood and *Wojood*$_{Fine}$. Note that creating *Wojood*$_{Fine}$ was not a straightforward process as it required revision of the Wojood annotation guidelines, which we discuss later in this section. As discussed in (Jarrar et al., 2022), Wojood is available as a RESTful web service, the data and the source-code are also made publicly available (Jarrar and Amayreh, 2019; Ghanem et al., 2023; Jarrar et al., 2019; Alhafi et al., 2019; Helou et al., 2016).

Tag	Wojood	*Wojood*$_{Fine}$
GPE	21,780	23,085
ORG	18,785	18,747
LOC	917	1,441
FAC	1,215	1,121
Total	**42,697**	**44,394**

Table 1: Frequency of the four entity types in Wojood and *Wojood*$_{Fine}$.

3.1 subtypes

All GPE, ORG, LOC and FAC tagged tokens in *Wojood*$_{Fine}$ corpus were annotated with the appropriate subtype based on the context, adding an additional 31 entity subtypes to *Wojood*$_{Fine}$. Throughout our annotation process, The LDC's ACE 2008 annotation guidelines for Arabic Entities V7.4.2 served as the basis for defining our annotation guidelines. Nevertheless, we added new tags (NEIGHBORHOOD, CAMP, SPORT, and ORG_FAC) to cover additional cases. Table 2 lists the frequency of each subtype in *Wojood*$_{Fine}$. Tables 7 and 8 in Appendix A present a brief explanation and examples of each subtype.

Tag	Sub-type Tag	Count
GPE	COUNTRY	8,205
	STATE-OR-PROVINCE	1,890
	TOWN	12,014
	NEIGHBORHOOD	119
	CAMP	838
	GPE_ORG	1,530
	SPORT	8
LOC	CONTINENT	214
	CLUSTER	303
	ADDRESS	0
	BOUNDARY	22
	CELESTIAL	4
	WATER-BODY	123
	LAND-REGION-NATURAL	259
	REGION-GENERAL	383
	REGION-INTERNATIONAL	110
ORG	GOV	8,325
	COM	611
	EDU	1,159
	ENT	3
	NONGOV	5,779
	MED	4,111
	REL	96
	SCI	146
	SPO	21
	ORG_FAC	114
FAC	PLANT	1
	AIRPORT	6
	BUILDING-OR-GROUNDS	1017
	SUBAREA-FACILITY	134
	PATH	76
Total		**47,621**

Table 2: Counts of each subtype entity in the corpus.

3.2 *Wojood*$_{Fine}$ Annotation Guideline

We followed ACE annotation guidelines to annotate the subtypes in *Wojood*$_{Fine}$. However, since *Wojood*$_{Fine}$ is based on Wojood, we found a discrepancy between Wojood and ACE guidelines. To address this issue in *Wojood*$_{Fine}$, we reviewed the annotations related to GPE, ORG, LOC and FAC to ensure compatibility with ACE guidelines. In this section, we highlight a number of the challenging annotation decisions we made in *Wojood*$_{Fine}$.

Country's governing body: in Wojood, country mentions were annotated as GPE and if the intended meaning of the country is a governing body then it is annotated as ORG. However, in *Wojood*$_{Fine}$, all ORG mentions that refer to the country's governing body are annotated as GPE with the subtype GPE_ORG. Figure 1 illustrates two examples to illustrate the difference between Wojood and *Wojood*$_{Fine}$ guidelines. According to Wojood, نيجيريا /Nigeria is tagged once as GPE and once as ORG, while in *Wojood*$_{Fine}$ both are GPE in the first level and in the second level one is tagged as Country and the other as GPE_ORG.

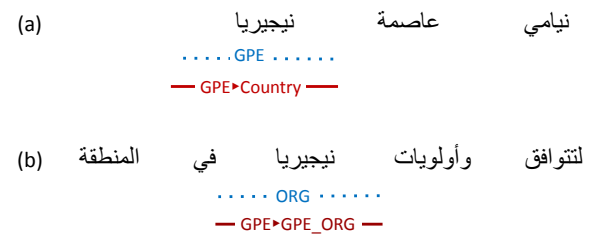

Figure 1: Two examples illustrating the difference between Wojood (in blue) and *Wojood*$_{Fine}$ guidelines (in red) for annotating GPEs.

Facility vs. organization: Wojood annotates buildings as FAC but if the intended meaning, in the context is an organization, then it is annotated as ORG. In *Wojood*$_{Fine}$, all mentions that refer to the facility's organization or social entity are annotated as ORG with the subtype ORG_FAC. Figure 2 illustrates an example of this case. Instead of annotating (مستشفى الشفاء /Al-Shifa Hospital) once as FAC and once as ORG, *Wojood*$_{Fine}$ tags it as ORG in the first level, and ORG_FAC in the second level.

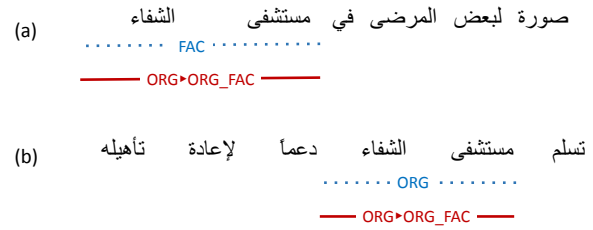

Figure 2: Two examples illustrating the difference between Wojood (in blue) and *Wojood*$_{Fine}$ (in red) guideline for annotating FAC vs. ORG.

Directions: Wojood does not include annotations for directions (east, west, south, and north). However, in *Wojood*$_{Fine}$ direction mentions are annotated as LOC with two subtypes: REGION-GENERAL if the location does not cross national borders, or REGION-INTERNATIONAL if the location crosses national borders. See the example in Figure 3.

In addition to the changes mentioned in this section, ACE guidelines considered any unit that is smaller-size than a village, like neighborhoods or camps, as LOC, while it is considered as GPE in Wojood guidelines. Continents are labaled as LOC in Wojood, while it is GPE in ACE. Both of these cases where corrected in *Wojood*$_{Fine}$.

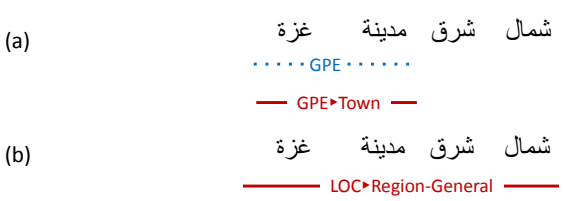

Figure 3: (a) The direction (شمال شرق مدينة غزة / north east Gaza city) is not annotated in Wojood, while in (b) it is annotated as LOC with Region-General as subtype in *Wojood*$_{Fine}$.

3.3 Annotation Process

The annotation process was done by one annotator, managed by NER expert, and was conducted over two phases:

Phase I: manually revise all annotations of GPE, ORG, LOC, and FAC in Wojood according to ACE guidelines, as discussed in section 3.2. Table 3.3 shows the counts of each of the four entity types in Wojood and *Wojood*$_{Fine}$.

Phase II: manually annotate the GPE, ORG, LOC, and FAC with subtypes. The annotator meticulously read each token in every sentence and classified the tokens into their respective subtypes. All critical and problematic tokens are reviewed by the NER expert.

Phase III: The NER expert reviewed all annotations marked in Phase I and Phase II in order to validate the entities that have been annotated.

Table 2 presents the counts of each entity subtype in the corpus, which shows 47,621 annotated entities in total.

3.4 Inter-Annotator Agreement

It has been shown that inter-annotator consistency significantly affects the quality of training data and, consequently, a NER system's ability to learn (Zhang, 2013). To measure the subtypes annotation quality and consistency, we recruited a second annotator to re-annotate 25,490 tokens (5.0% of the corpus) that were previously annotated by the first annotator. The sentences were selected randomly from the corpus while diversifying the sources and domains they were selected from. We then assessed the data quality and annotation consistency using the inter-annotator agreement (IAA), measured using Cohen's Kappa (κ) and F_1. The overall IAA was measured at $\kappa = 0.9861$ and $F_1 = 0.9889$.

Refer to Table 3 for the IAA for each subtype.

One can clearly observe that κ is high and that is for multiple reasons. First, we revised the annotations of the main four entity types (GPE, ORG, LOC and FAC) to better match ACE guideline. Second, once we verified the top level entity types, we started annotating the subtypes. Since the types and subtypes are hierarchically organized, that constraint the number of possible subtypes per token, leading to high IAA. Third, the NER expert gave a continuous feedback to the annotator and challenging entity mentions were discussed with the greater team.

As mentioned above, we calculated the IAA using both, Cohen's Kappa and F_1, for the subtypes of GPE, ORG, LOC and FAC tags. In what follows we explain Cohen's Kappa and F_1. Note that F_1 is not normally used for IAA, but it is an additional validation of the annotation quality.

3.4.1 Cohen's Kappa

To calculate Kappa for a given tag, we count the number of agreements and disagreements between annotators for a given subtype (such as GPE_COUNTRY). At the token level, agreements are counted as pairwise matches; thus, disagreements happen when a token is annotated by one annotator (e.g., as GPE_COUNTRY) and (e.g., as GPE_STATE-OR-PROVINCE) by another annotator. As such, Kappa is calculated by equation 1 (Eugenio and Glass, 2004).

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{1}$$

where P_o represents the observed agreement between annotators and P_e represents the expected agreement, which is given by equation 2.

$$P_e = \frac{1}{N^2} \sum_{T} n_{T1} \times n_{T2} \tag{2}$$

where n_{Ti} is the number of tokens labeled with tag T by the ith annotator and N is the total number of annotated tokens.

3.4.2 F-Measure

For a given tag T, the F_1 is calculated according to equation 3. We only counted the tokens that at least one of the annotators had labeled with the T. We then conducted a pair-wise comparison. TP represents the true positives which is the number of agreements between annotators (i.e. number of tokens labeled GPE_TOWN by both annotators). If the first annotator disagrees with the second, it is counted as false negatives (FN), and if the second disagrees with the first, it is counted as false positives (FP), with a total of disagreement being $FN + FP$.

$$F_1 = \frac{2TP}{2TP + FN + FP} \tag{3}$$

Sub-Type Tag	Kappa	F1-Score
COUNTRY	0.9907	00.99
STATE-OR-PRONIVCE	0.9846	00.98
TOWN	0.9983	01.00
NEIGHBORHOOD	01.00	01.00
CAMP	01.00	01.00
GPE_ORG	0.9810	00.98
SPORT	01.00	01.00
CONTINENT	01.00	01.00
CLUSTER	0.9589	00.96
ADDRESS	-	-
BOUNDARY	01.00	01.00
CELESTIAL	-	-
WATER-BODY	01.00	01.00
LAND-REGION-NATURAL	0.9333	00.93
REGION-GENERAL	0.9589	00.96
REGION-INTERNATIONAL	0.9231	00.92
GOV	0.9760	00.98
COM	01.00	01.00
EDU	0.9807	00.98
ENT	-	-
NONGOV	0.9892	00.99
MED	01.00	01.00
REL	0.9630	00.96
SCI	01.00	00.10
SPO	01.00	01.00
ORG_FAC	01.00	01.00
PLANT	-	-
AIRPORT	-	-
BUILDING-OR-GROUNDS	01.00	01.00
SUBAREA-FACILITY	01.00	01.00
PATH	01.00	00.00
Overall	**0.9861**	**0.9889**

Table 3: Overall Kappa and F1-score for each sub-type.

4 Fine-Grained NER Modeling

4.1 Approach

For modeling, we have three tasks all performed on *Wojood*$_{Fine}$: **(1)** *Flat NER*, where for each token, we predict a single label from a set of 21 labels, **(2)** *Nested NER*, where we predict multiple labels picked from the 21 tags (i.e., multi-label classification) for each token and **(3)** *Nested with Subtypes NER*, this is also a multi-label task, where we ask the model to predict the main entity types and subtypes for each token from 52 total labels. We frame this as multi-task approach

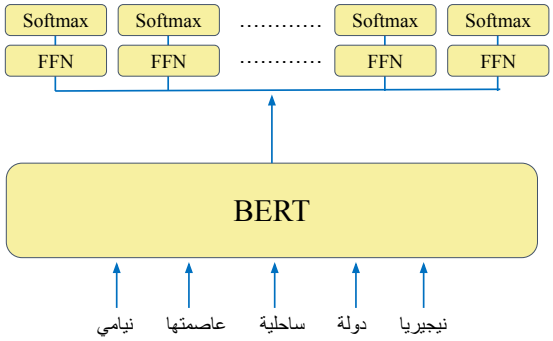

Figure 4: BERT refers to one of three pre-trained models we are using. For flat task, each softmax produce one class for each token, for other tasks each softmax is a set of softmax that produce multiple labels for each token.

since we are learning both the nested labels *and* their subtypes jointly. In the multi-task case, each entity/subtype has its own classification layer, in the case of nested NER and nested with subtypes NER, the model consists of 21 and 52 classification layers, respectively. Since we use the IOB2 (Sang and Veenstra, 1999) tagging scheme, each linear layer is a multi-class classifier that outputs the probability distribution through softmax activation function for three classes, $C \in \{I, O, B\}$ (Jarrar et al., 2022). The model is trained with cross entropy loss objective computed for each linear layer separately, which are summed to compute the final cross entropy loss. All models are flat in the sense that we do not use any hierarchical architectures. However, future work can consider employing a hierarchical architecture where nested tokens are learnt first *then* their subtypes within the model. For all tasks, we fine-tune three encoder-based models for Arabic language understanding. Namely, we use ARBERTv2 and MARBERTv2 (Elmadany et al., 2023), which are both improved versions of ARBERT and MARBERT (Abdul-Mageed et al., 2021), respectively, that are trained on bigger datasets. The third model is ARABERTv2, which is an improved version of ARABERT (Antoun et al., 2021). It is also trained on a bigger dataset, with improved preprocessing. Figure 4 offers a simple visualization of our models' architecture.

4.2 Training Configuration

We split our dataset into three distinct parts for training (Train) 70%, validation (Dev) 10%, and blind testing (Test) 20%. We fine-tune all three models for 50 epochs each with an early stopping patience of 5 as identified on Dev. We use the AdamW optimizer (Loshchilov and Hutter, 2019), an exponential learning rate scheduler and a dropout of 0.1. The maximum sequence length is 512, the batch size, $B = 8$, and the learning rate, $\eta = 1e^{-5}$. For each model, we report an average of three runs (each time with a different seed). We report in F_1 along with the standard deviation from the three runs, on both Dev and Test, for each model. All models are implemented using PyTorch, Huggingface Transformers, and a custom version of the Wojood open-source code[1].

Task	Model	Dev	Test
Flat	M1	$0.917^{\pm 0.00}$	$\mathbf{0.920}^{\pm 0.00}$
	M2	$0.910^{\pm 0.00}$	$0.913^{\pm 0.01}$
	M3	$0.902^{\pm 0.00}$	$0.907^{\pm 0.01}$
Nested	M1	$0.844^{\pm 0.02}$	$0.845^{\pm 0.01}$
	M2	$0.868^{\pm 0.02}$	$0.861^{\pm 0.02}$
	M3	$0.858^{\pm 0.02}$	$\mathbf{0.866}^{\pm 0.02}$
Nested +subtypes	M1	$0.836^{\pm 0.01}$	$0.837^{\pm 0.01}$
	M2	$0.880^{\pm 0.01}$	$0.883^{\pm 0.01}$
	M3	$0.883^{\pm 0.00}$	$\mathbf{0.885}^{\pm 0.00}$

Table 4: Results of fine-tuned models on the three different tasks. **M1**: ARBERTv2, **M2**: MARBERTv2 and **M3**: ARABERTv2. The results are represented as F1 averaged over 3 runs.

4.3 Results

We show the results of our three fine-tuned models across each of the three tasks in Table 4. We briefly highlight these results in the following:

Flat NER. The three fine-tuned models achieve comparable results on the Flat NER task, with ARBERTv2 scoring slightly better on both the Dev and Test sets. ARBERTv2 achieves an F_1 of 92% on the Test set, while ARBERTv2 and ARABERTv2 achieves 91.3% and 90.3%, respectively.

Nested NER. ARABERTv2 slightly outperforms other pre-trained models with a small margin, on Dev and Test. On Test, it scores 86.6%.

Nested NER with Subtypes. Here, ARABERTv2 achieves the highest score (88.5%F_1).

5 Analysis

For all tasks, all models almost always converge in the first 10 epochs. For all models, there is a positive correlation between performance and the number of training samples. For example, for classes represented well in the training set (e.g.,

[1] https://github.com/SinaLab/ArabicNER

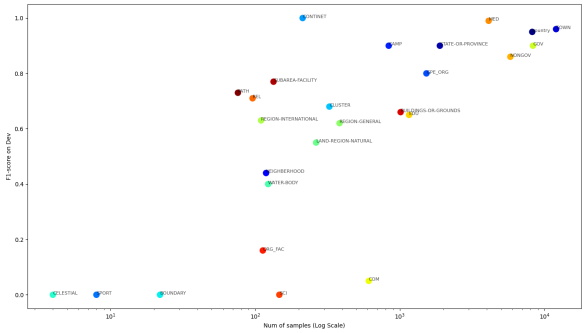

Figure 5: Number of samples vs. F_1 in each subtype class on Subtype classification task.

Task	Model	Finance	Science	Politics
Flat	M1	63.7% $^{\pm 0.01}$	0.670$^{\pm 0.02}$	**0.747**$^{\pm 0.02}$
	M2	0.573$^{\pm 0.01}$	**0.677**$^{\pm 0.02}$	0.717$^{\pm 0.01}$
	M3	**0.643**$^{\pm 0.01}$	0.670$^{\pm 0.02}$	0.723$^{\pm 0.01}$
Nested	M1	0.458$^{\pm 0.01}$	0.494$^{\pm 0.02}$	0.557$^{\pm 0.00}$
	M2	0.499$^{\pm 0.05}$	0.554$^{\pm 0.00}$	0.612$^{\pm 0.01}$
	M3	**0.563**$^{\pm 0.02}$	**0.583**$^{\pm 0.02}$	**0.629**$^{\pm 0.03}$
Nested +subtypes	M1	0.449$^{\pm 0.07}$	0.493$^{\pm 0.02}$	0.497$^{\pm 0.01}$
	M2	0.504$^{\pm 0.03}$	0.544$^{\pm 0.06}$	0.575$^{\pm 0.02}$
	M3	**0.553**$^{\pm 0.04}$	**0.545**$^{\pm 0.02}$	**0.593**$^{\pm 0.08}$

Table 5: Results of fine-tuned models on the three new domains, Finance, Science, and Politics. **M1**: MARBERTv2, **M2**: ARBERTv2 and **M3**: ARABERTv2. The results are represented as F1 averaged over 3 runs.

COUNTRY, TOWN and GOV), models perform at 0.90 F_1 or above.

The inverse is also true, with poor performance on classes such as SPORT, BOUNDARY and CELESTIAL. There are also some nuances. For example, we can see that the best model is struggling with the COM subtype class even though the model has scored good results with classes with fewer samples such as CLUSTER. The main reason for this is that types such as CLUSTER are a closed set of classes (e.g., "European Union", "African Union") where the model can easily memorize them, while the COM refers to an infinite group of commercial entities, that can not be limited. Figure 5 is a plot of the number of samples in training data (X-axis) vs. performance (Y-axis) that clearly shows the general pattern of good performance positively correlating with the number of training samples.

5.1 Out-of-Domain Performance

To assess the generalization capability of our models, we conducted an evaluation on three unseen domains and different time periods. Three corpora were collected, each covering a distinct domain: finance, science, and politics. These corpora were compiled from Aljazeera news articles published in 2023. Manual annotation of the three corpora was performed in accordance with the same annotation guidelines established for *Wojood*$_{Fine}$. We apply the three versions of each of our three models trained on *Wojood*$_{Fine}$ original training data (described in Section 4.2) on the new domains, for each of the three NER tasks. We present results for this out-of-domain set of experiments in Table 5. We observe that performance drastically drops on all three new domains, for all models on all tasks. This is not surprising, as challenges related to domain generalization are well-known in the literature. Our results here, however, allow us to quantify the extent to which model performance degrades on each of these three new domains. In particular, models do much better on the politics domain than they perform on finance or science. This is the case since our training data are collected from online articles involving news and much less content from financial or scientific sources. Figure 6 shows some examples for new mentions from those domains that have not been seen in *Wojood*$_{Fine}$.

Figure 6: Some mentions from the three new domains that have not previously appeared in *Wojood*$_{Fine}$. (a) (مركز المعلومات الفلسطيني) in Politics domain, (b) (مجموعة إنتل) in Finance domain, (c) (منظمة OpenAI) in Science domain.

5.2 Error Analysis

In order to understand the errors made by the model, we conduct a human error analysis on the errors generated by ARABERTv2 (i.e, best model on this task) on the first 2K tokens of the Dev set of Nested NER with Subtypes task. We find that the model's errors can be categorized into six major error classes: **(1)** *wrong tag*, where the model predicts a different tag, **(2)** *no prediction*, where the model does not produce any tag (i.e. predict O), **(3)** *missing subtype*, the model succeeds in predicting parent tag but fails to predict the subtype,

Example	Gold	Predicted	Error Type
أنا ازا هاجرت ع أي مكان رح آخد الشلة If I ever migrated somewhere, I'd take the group	O	GPE|TWN	msa_dia_confusion
مشهد ٣ فتاة جالسة و خلفها العلم الأمريكي. Scene 3: a girl sitting with the American flag behind her.	CRDNAL	ORDNAL	ordinal_vs_cardinal
جدار الفصل العنصري مستعمرة بزغات زئيف. The racial separation wall, colony of Bazgat Ze'ev.	LOC|NEIGHB	NEIGHB	Missing_parent_tag
بتنتخب رئيس جمهورية و رئيس مجلس نواب The president of the republic and the speaker of the council of deputies are elected.	OCC|ORG|GOV	OCC|ORG	missing_subtype
صحيح الساعة خمسة حسب اعلانهم It's true, it's five o'clock according to their announcement.	TIME	CRDNL	wrong_tag
العلما اللغة التانية بتنحصر للاستخدام اليومي. Scientists: the second language is limited to daily use.	B-ORDNL	O	no_prediction

Table 6: Examples of error categories made by our best model (ARABERTv2) on our Dev set. We provide the translation to English of each sample.

(4) *missing parent tag*: the model succeeds in predicting subtype tag but fails to predict the parent tag, **(5)** *MSA vs. DIA confusion*, the model makes a wrong prediction due to confusion between MSA and Dialect, and **(6)** *ordinal vs. cardinal*, in this class, the model assigns cardinal to an ordinal class. Figure 7 shows the distribution of different errors present in the Dev set, with the *wrong tag* being the major source of errors followed by *no prediction* error. A further breakdown of the *wrong tag* error class shows that 14.3% are due to usage of dialectal words, a similar proportion are due to nested entities. Table 6 shows an example of each error class.

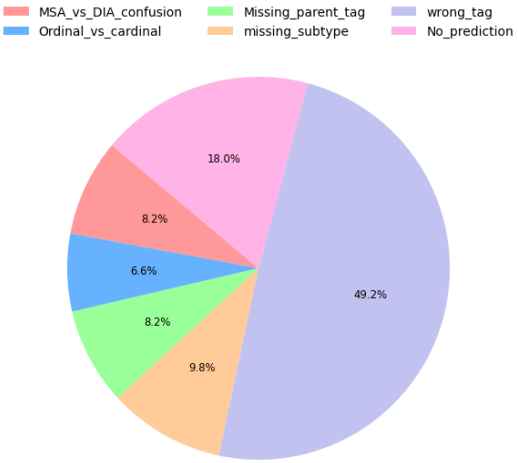

Figure 7: Distribution of error classes in nested with subtypes task on our Dev set.

6 Conclusion and Future Work

We presented *Wojood*$_{Fine}$, an extension to the Wojood NER corpus with subtypes for the GPE, LOC, ORG, and FAC. *Wojood*$_{Fine}$ corpus is the first fine-grain corpus for MSA and dialectal Arabic with nested and subtyped NER. The GPE, ORG, FAC and LOC tags form more than 44K tokens of the corpus, which was manually annotated using subtypes entities. Our inter-annotator agreement IAA evaluation of *Wojood*$_{Fine}$ annotations achieved high levels of agreement among the annotators. The achieved evaluations are 0.9861 Kappa and 0.9889 F_1.

We also fine-tune three pre-trained models ARBERTv2, MARBERTv2 and ARABERTv2 and tested their performance on different settings of *Wojood*$_{Fine}$. We find that ARABERTv2 achieved the best performance on Nested and Nested with Subtypes tasks. In the future, we plan to test pre-trained models on nested subtypes with hierarchical architecture. We also plan to link named entities with concepts in the Arabic Ontology (Jarrar, 2021, 2011) to enable a richer semantic understanding of text. Additionally, we will extend the *Wojood*$_{Fine}$ corpus to include more dialects, especially the Syrian Nabra dialects (Nayouf et al., 2023) as well as the four dialects in the Lisan (Jarrar et al., 2023b) corpus.

Acknowledgment

We would like to thank Sana Ghanem for her invaluable assistance in reviewing and improving the annotations, as well as for her support in the IAA calculations. The authors would also like to thank Tymaa Hammouda for her technical support and

expertise in the data engineering of the corpus.

Limitations

A number of considerations related to limitations and ethics are relevant to our work, as follows:

- **Intended Use.** Our models perform named entity recognition at a fine-grained level and can be used for a wide range of information extraction tasks. As we have shown, however, even though the models are trained with data acquired from several domains, their performance drops on data with distribution different than our training data such as the finance or science domains. We suggest this be taken into account in any application of the models.
- **Annotation Guidelines and Process.** Some of the entities are difficult to tag. Even though annotators have done their best and we report high inter-annotator reliability, the application of our guidelines may need to be adapted before application to new domains.

Ethics Statement

We trained our models on publicly available data, thus we do not have any particular concerns about privacy.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *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)*, pages 7088–7105, Online. Association for Computational Linguistics.

Moustafa Al-Hajj and Mustafa Jarrar. 2021. Arabglossbert: Fine-tuning bert on context-gloss pairs for wsd. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 40–48, Online. INCOMA Ltd.

Diana Alhafi, Anton Deik, and Mustafa Jarrar. 2019. Usability evaluation of lexicographic e-services. In *The 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*, pages 1–7. IEE.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. Arabert: Transformer-based model for arabic language understanding.

Yassine Benajiba, Paolo Rosso, and José Miguel Benedi Ruiz. 2007. Anersys: An arabic named entity recognition system based on maximum entropy. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 4394 LNCS.

Nancy Chinchor and Patricia Robinson. 1997. Muc-7 named entity task definition. In *Proceedings of the 7th Conference on Message Understanding*, volume 29, pages 1–21.

Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein T. Al-Natsheh, Houda Bouamor, Karim Bouzoubaa, Violetta Cavalli-Sforza, Samhaa R. El-Beltagy, Wassim El-Hajj, Mustafa Jarrar, and Hamdy Mubarak. 2021. A panoramic survey of natural language processing in the arab worlds. *Commun. ACM*, 64(4):72–81.

Bart Desmet and Véronique Hoste. 2013. Fine-grained dutch named entity recognition. *Language Resources and Evaluation*, 48:307–343.

Maud Ehrmann, Matteo Romanello, Alex Flückiger, and Simon Clematide. 2020a. Extended overview of clef hipe 2020: named entity processing on historical newspapers. In *CLEF 2020 Working Notes. Conference and Labs of the Evaluation Forum*, volume 2696. CEUR-WS.

Maud Ehrmann, Camille Watter, Matteo Romanello, Clematide Simon, and Alex Flückiger. 2020b. Impresso named entity annotation guidelines (clef-hipe-2020). Technical report.

AbdelRahim Elmadany, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Orca: A challenging benchmark for arabic language understanding.

Barbara Di Eugenio and Michael Glass. 2004. The Kappa Statistic: A Second Look. *Computational Linguistics*, 30(1):95–101.

Besnik Fetahu, Sudipta Kar, Zhiyu Chen, Oleg Rokhlenko, and Shervin Malmasi. 2023. Semeval-2023 task 2: Fine-grained multilingual named entity recognition (multiconer 2). *arXiv preprint arXiv:2305.06586*.

Sana Ghanem, Mustafa Jarrar, Radi Jarrar, and Ibrahim Bounhas. 2023. A benchmark and scoring algorithm for enriching arabic synonyms. In *Proceedings of the 12th International Global Wordnet Conference (GWC2023)*, pages 215–222. Global Wordnet Association.

Karim El Haff, Mustafa Jarrar, Tymaa Hammouda, and Fadi Zaraket. 2022. Curras + baladi: Towards a levantine corpus. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2022)*, Marseille, France.

Ahmed Hamdi, Elvys Linhares Pontes, Emanuela Boros, Thi Tuyet Hai Nguyen, Günter Hackl, Jose G Moreno, and Antoine Doucet. 2021. A multilingual dataset for

named entity recognition, entity linking and stance detection in historical newspapers. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2328–2334.

Mamoun Abu Helou, Matteo Palmonari, and Mustafa Jarrar. 2016. Effectiveness of automatic translations for cross-lingual ontology mapping. *Journal of Artificial Intelligence Research*, 55(1):165–208.

Alankar Jain, Bhargavi Paranjape, and Zachary C Lipton. 2019. Entity projection via machine translation for cross-lingual ner. *arXiv preprint arXiv:1909.05356*.

Mustafa Jarrar. 2011. Building a formal arabic ontology (invited paper). In *Proceedings of the Experts Meeting on Arabic Ontologies and Semantic Networks*. ALECSO, Arab League.

Mustafa Jarrar. 2021. The arabic ontology - an arabic wordnet with ontologically clean content. *Applied Ontology Journal*, 16(1):1–26.

Mustafa Jarrar and Hamzeh Amayreh. 2019. An arabic-multilingual database with a lexicographic search engine. In *The 24th International Conference on Applications of Natural Language to Information Systems (NLDB 2019)*, volume 11608 of *LNCS*, pages 234–246. Springer.

Mustafa Jarrar, Hamzeh Amayreh, and John P. McCrae. 2019. Representing arabic lexicons in lemon - a preliminary study. In *The 2nd Conference on Language, Data and Knowledge (LDK 2019)*, volume 2402, pages 29–33. CEUR Workshop Proceedings.

Mustafa Jarrar, Anton Deik, and Bilal Faraj. 2011. Ontology-based data and process governance framework -the case of e-government interoperability in palestine. In *Proceedings of the IFIP International Symposium on Data-Driven Process Discovery and Analysis (SIMPDA'11)*, pages 83–98.

Mustafa Jarrar, Nizar Habash, Diyam Akra, and Nasser Zalmout. 2014. Building a corpus for palestinian arabic: a preliminary study. In *Proceedings of the EMNLP 2014, Workshop on Arabic Natural Language*, pages 18–27. Association For Computational Linguistics.

Mustafa Jarrar, Nizar Habash, Faeq Alrimawi, Diyam Akra, and Nasser Zalmout. 2017. Curras: An annotated corpus for the palestinian arabic dialect. *Journal Language Resources and Evaluation*, 51(3):745–775.

Mustafa Jarrar, Mohammed Khalilia, and Sana Ghanem. 2022. Wojood: Nested arabic named entity corpus and recognition using bert. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2022)*, Marseille, France.

Mustafa Jarrar, Sanad Malaysha, Tymaa Hammouda, and Mohammad Khalilia. 2023a. Salma: Arabic sense-annotated corpus and wsd benchmarks. In *Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023*. ACL.

Mustafa Jarrar, Fadi Zaraket, Tymaa Hammouda, Daanish Masood Alavi, and Martin Waahlisch. 2023b. Lisan: Yemeni, irqi, libyan, and sudanese arabic dialect copora with morphological annotations. In *The 20th IEEE/ACS International Conference on Computer Systems and Applications (AICCSA)*. IEEE.

Ridong Jiang, Rafael E. Banchs, and Haizhou Li. 2016. Evaluating and combining name entity recognition systems. In *Proceedings of the Sixth Named Entity Workshop*, pages 21–27, Berlin, Germany. Association for Computational Linguistics.

Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. 2022. Natural language processing: State of the art, current trends and challenges. *Multimedia Tools and Applications*, 82.

Changki Lee, Yi-Gyu Hwang, Hyo-Jung Oh, Soojong Lim, Jeong Heo, Chung-Hee Lee, Hyeon-Jin Kim, Ji-Hyun Wang, and Myung-Gil Jang. 2006. Fine-grained named entity recognition using conditional random fields for question answering. In *Information Retrieval Technology*, pages 581–587, Berlin, Heidelberg. Springer Berlin Heidelberg.

Jing li, Aixin Sun, Ray Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, PP:1–1.

Xiao Ling and Daniel Weld. 2012. Fine-grained entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pages 94–100.

Bang Liu, Haojie Wei, Di Niu, Haolan Chen, and Yancheng He. 2020. Asking questions the human way: Scalable question-answer generation from text corpus. In *Proceedings of The Web Conference 2020*, pages 2032–2043.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Amal Nayouf, Mustafa Jarrar, Fadi zaraket, Tymaa Hammouda, and Mohamad-Bassam Kurdy. 2023. Nâbra: Syrian arabic dialects with morphological annotations. In *Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023*. ACL.

Ramzi Esmail Salah and Lailatul Qadri Binti Zakaria. 2018. Building the classical arabic named entity recognition corpus (canercorpus). *Journal of Theoretical and Applied Information Technology*, 96.

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Erik F. Tjong Kim Sang and Jorn Veenstra. 1999. Representing text chunks. In *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics*, EACL '99, page 173–179, USA. Association for Computational Linguistics.

Rodney L Summerscales, Shlomo Argamon, Shangda Bai, Jordan Hupert, and Alan Schwartz. 2011. Automatic summarization of results from clinical trials. In *2011 IEEE International Conference on Bioinformatics and Biomedicine*, pages 372–377. IEEE.

Mikhail Tikhomirov, N. Loukachevitch, Anastasiia Sirotina, and Boris Dobrov. 2020. Using bert and augmentation in named entity recognition for cybersecurity domain. In *Natural Language Processing and Information Systems*, pages 16–24, Cham. Springer International Publishing.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2005. Ace 2005 multilingual training corpus-linguistic data consortium. *URL: https://catalog. ldc. upenn. edu/LDC2006T06*.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0 ldc2013t19. Technical report, Linguistic Data Consortium.

Ziqi Zhang. 2013. Named entity recognition : challenges in document annotation, gazetteer construction and disambiguation.

Huiming Zhu, Chunhui He, Yang Fang, and Weidong Xiao. 2020. Fine grained named entity recognition via seq2seq framework. *IEEE Access*, 8:53953–53961.

A subtypes and Inter-Annotator Agreement

This is an appendix contains *Wojood*$_{Fine}$ subtype descriptions and detailed IAA.

Tag	Sub-type Tag	Short Description
GPE	COUNTRY	Taggable mentions of the entireties of any nation. فلسطين، مصر، الولايات المتحدة، لبنان.
	STATE-OR-PRONIVCE	Taggable mentions of the entireties of any state, province, or canton of any nation. إقليم كردستان، لواء نابلس.محافظة القاهرة، قطاع غزة،
	TOWN	Taggable mentions of any GPE entireties below the level of State-or-Province, including cities, and villages. العاصمة دبي، قرية بيرزيت.
	NEIGHBORHOOD	Taggable mentions of the entireties of units that are smaller than villages. حي الطيرة، البلدة القديمة، حي المغاربة.
	CAMP	Taggable mentions of the entireties of units that are smaller than villages, relating to refugees. مخيم قلنديا، مخيم نور شمس.
	GPE_ORG	is used for GPE mentions that refer to the entire governing body of a GPE. قررت فلسطين إعفاء المتضررين.أصدرت الولايات المتحدة تقريرها،
	SPORT	Athletes, Sports Teams. برشلونة، ميلان. مباراة المغرب، الفرق الرياضية.
LOC	CONTINENT	Taggable mentions of the entireties of any of the seven continents. أوروبا، آسيا.
	CLUSTER	Named groupings of GPEs that can function as political entities. أوروبا الشرقية، الشرق الأوسط.
	ADDRESS	A location denoted as a point such as in a postal system ("31° S, 22° W"). ١٧، شارع فؤاد.
	BOUNDARY	A one-dimensional location such as a border between GPE's or other locations. الحدود الشرقية، الحدود السورية التركية.
	CELESTIAL	world, earth, globe in addition to all other planets. المريخ، عطارد.
	WATER-BODY	Bodies of water, natural or artificial (man-made). البحر الأحمر، الأطلسي.
	LAND-REGION-NATURAL	Geologically or ecosystemically designated, non-artificial locations. جبال الألب، الأغوار، السهول.
	REGION-GENERAL	Taggable locations that do not cross national borders. شمال الضفة الغربية، شرق سوريا.
	REGION-INTERNATIONAL	Taggable locations that cross national borders. آسيا الكبرى، جنوب أفريقيا.

Table 7: Parent type and description of each sub-type in *Wojood*$_{Fine}$

Tag	Sub-type Tag	Short Description
ORG	GOV	Government organizations. . سفارة، محكمة، وزارة، شرطة
	COM	A commercial organization that is focused primarily upon providing ideas, products, or services for profit. .بنك ، شركة ،مؤسسة ربحية
	EDU	An educational organization that is focused primarily upon the furthering or promulgation of learning/education. .جامعة، مدرسة، معهد
	ENT	Entertainment organizations whose primary activity is entertainment. فرقة ميامي، مسرح الحكواتي.
	NONGOV	Non-governmental organizations that are not a part of a government or commercial organization and whose main role is advocacy, charity or politics (in a broad sense). .نقابة العاملين، الأمم المتحدة، الأحزاب السياسية ،أطباء بلا حدود
	MED	Media organizations whose primary interest is the distribution of news or publications. .جريدة الشرق، مجلة الحياة
	REL	Religious organizations that are primarily devoted to issues of religious worship. .الأوقاف ، الأزهر
	SCI	Medical-Science organizations whose primary activity is the application of medical care or the pursuit of scientific research. .مستشفى هداسا ،معهد الدراسات النووية
	SPO	Sports organizations that are primarily concerned with participating in or governing organized sporting events. .الاتحاد السعودي لكرة القدم،لجنة الفلبين الأولومبية
	ORG_FAC	Facilities that have an organizational, legal or social representative مظاهرات أمام بنك روما.
FAC	PLANT	One or more buildings that are used and/or designed solely for industrial purposes: manufacturing, power generation, etc. .مصنع
	AIRPORT	A facility whose primary use is as an airport. .مطار
	BUILDING-OR-GROUNDS	Man-made/-maintained buildings, outdoor spaces, and other such facilities. .منزل، مبنى، مستشفى، معبر
	SUBAREA-FACILITY	Taggable portions of facilities. .غرفة ،زنزانة
	PATH	Streets, canals, and bridges. .الشوارع الرئيسية، الخطوط الهاتفية، الحواجز

Table 8: Parent type and description of each sub-type in *Wojood*$_{Fine}$

Sub-type Tag	TP	FN	FP	Kappa	F1-Score
COUNTRY	643	5	7	0.9907	00.99
STATE-OR-PRONIVCE	96	3	0	0.9846	00.98
TOWN	295	0	1	0.9983	01.00
NEIGHBORHOOD	23	0	0	01.00	01.00
CAMP	92	0	0	01.00	01.00
GPE_ORG	129	3	2	0.9810	00.98
SPORT	2	0	0	01.00	01.00
CONTINENT	7	0	0	01.00	01.00
CLUSTER	35	3	0	0.9589	00.96
ADDRESS	-	-	-	-	-
BOUNDARY	11	0	0	01.00	01.00
CELESTIAL	-	-	-	-	-
WATER-BODY	5	0	0	01.00	01.00
LAND-REGION-NATURAL	14	0	2	0.9333	00.93
REGION-GENERAL	70	2	4	0.9589	00.96
REGION-INTERNATIONAL	6	0	1	0.9231	00.92
GOV	490	6	18	0.9760	00.98
COM	21	0	0	01.00	01.00
EDU	153	0	6	0.9807	00.98
ENT	-	-	-	-	-
NONGOV	599	11	2	0.9892	00.99
MED	630	0	0	01.00	01.00
REL	26	2	0	0.9630	00.96
SCI	4	0	0	01.00	00.10
SPO	2	0	0	01.00	01.00
ORG_FAC	15	0	0	01.00	01.00
PLANT	-	-	-	-	-
AIRPORT	-	-	-	-	-
BUILDING-OR-GROUNDS	64	0	0	01.00	01.00
SUBAREA-FACILITY	48	0	0	01.00	01.00
PATH	2	0	0	01.00	01.00
Overall	**3,482** count	**35** count	**43** count	**0.9861** macro	**0.9889** micro

Table 9: Overall IAA for each sub-type, reported using Kappa and F_1.