The 2023 Conference on Empirical Methods in Natural Language Processing
Singapore, 7 December, 2023

WojoodNER 2023: The First Arabic Named Entity Recognition Shared Task

Mustafa Jarrar Muhammad Abdul-Mageed Mohammad Khalilia Bashar Talafha
AbdelRahim Elmadany Nagham Hamad Alaa' Omar

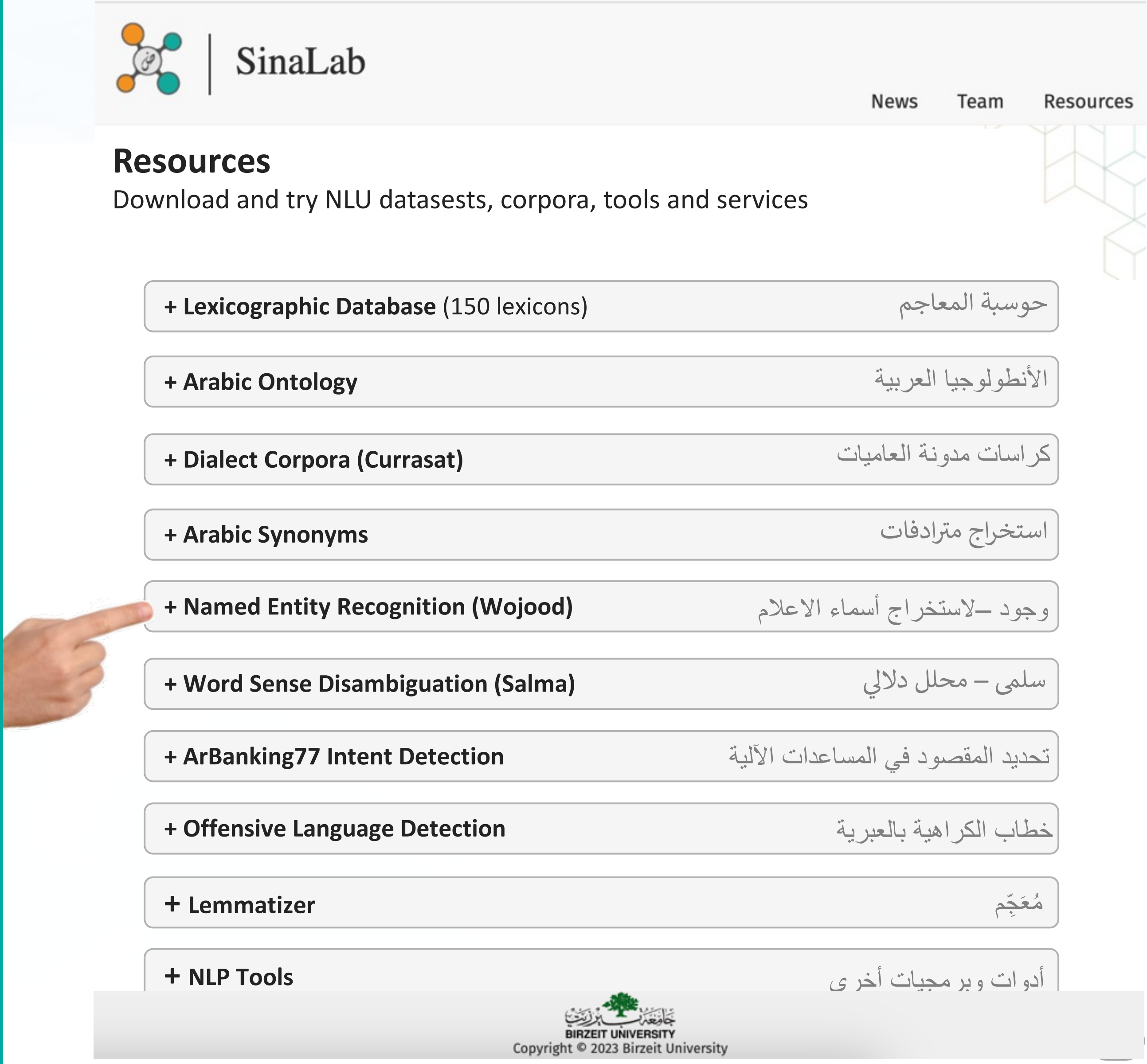
SinaLab, Birzeit University,
Palestine

The University of British Columbia
Canada

Natural Language Understanding Tools and Datasets

Open Source

https://sina.birzeit.edu/resources

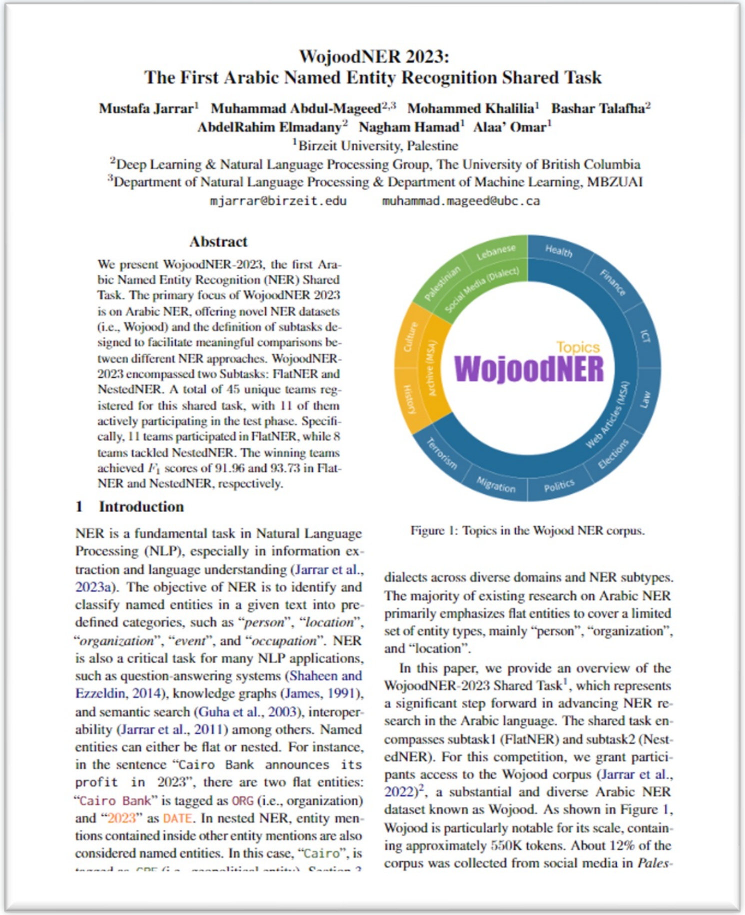


Resources

Download and try NLU datasests, corpora, tools and services

+ **Lexicographic Database** (150 lexicons)	حوسبة المعاجم
+ **Arabic Ontology**	الأنطولوجيا العربية
+ **Dialect Corpora (Currasat)**	كراسات مدونة العاميات
+ **Arabic Synonyms**	استخراج مترادفات
+ **Named Entity Recognition (Wojood)**	وجود –لاستخراج أسماء الاعلام
+ **Word Sense Disambiguation (Salma)**	سلمى – محلل دلالي
+ **ArBanking77 Intent Detection**	تحديد المقصود في المساعدات الآلية
+ **Offensive Language Detection**	خطاب الكراهية بالعبرية
+ **Lemmatizer**	مُعَجِّم
+ **NLP Tools**	أدوات وبرمجيات أخرى

جامعة بيرزيت
BIRZEIT UNIVERSITY
Copyright © 2023 Birzeit University



WojoodNER 2023:
The First Arabic Named Entity Recognition Shared Task

Mustafa Jarrar[1] Muhammad Abdul-Mageed[2,3] Mohammed Khalilia[1] Bashar Talafha[2]
AbdelRahim Elmadany[2] Nagham Hamad[1] Alaa' Omar[1]
[1]Birzeit University, Palestine
[2]Deep Learning & Natural Language Processing Group, The University of British Columbia
[3]Department of Natural Language Processing & Department of Machine Learning, MBZUAI
mjarrar@birzeit.edu muhammad.mageed@ubc.ca

Abstract

We present WojoodNER-2023, the first Arabic Named Entity Recognition (NER) Shared Task. The primary focus of WojoodNER 2023 is on Arabic NER, offering novel NER datasets (i.e., Wojood) and the definition of subtasks designed to facilitate meaningful comparisons between different NER approaches. WojoodNER-2023 encompassed two Subtasks: FlatNER and NestedNER. A total of 45 unique teams registered for this shared task, with 11 of them actively participating in the test phase. Specifically, 11 teams participated in FlatNER, while 8 teams tackled NestedNER. The winning teams achieved F_1 scores of 91.96 and 93.73 in FlatNER and NestedNER, respectively.

1 Introduction

NER is a fundamental task in Natural Language Processing (NLP), especially in information extraction and language understanding (Jarrar et al., 2023a). The objective of NER is to identify and classify named entities in a given text into predefined categories, such as "*person*", "*location*", "*organization*", "*event*", and "*occupation*". NER is also a critical task for many NLP applications, such as question-answering systems (Shaheen and Ezzeldin, 2014), knowledge graphs (James, 1991), and semantic search (Guha et al., 2003), interoperability (Jarrar et al., 2011) among others. Named entities can either be flat or nested. For instance, in the sentence "Cairo Bank announces its profit in 2023", there are two flat entities: "Cairo Bank" is tagged as ORG (i.e., organization) and "2023" as DATE. In nested NER, entity mentions contained inside other entity mentions are also considered named entities. In this case, "Cairo", is tagged as GPE (i.e., geopolitical entity). Section 3

Figure 1: Topics in the Wojood NER corpus.

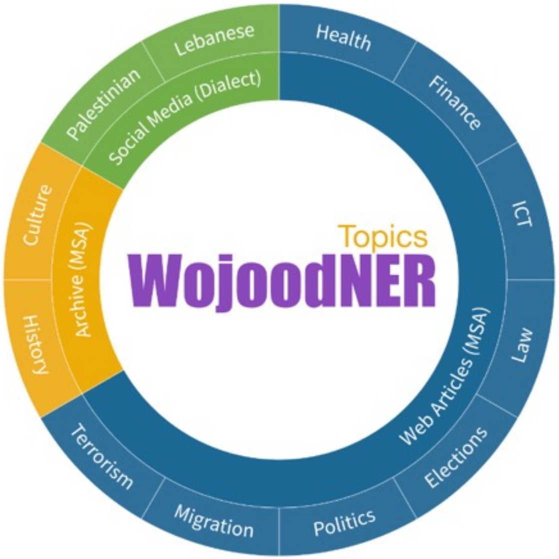
dialects across diverse domains and NER subtypes. The majority of existing research on Arabic NER primarily emphasizes flat entities to cover a limited set of entity types, mainly "person", "organization", and "location".

In this paper, we provide an overview of the WojoodNER-2023 Shared Task[1], which represents a significant step forward in advancing NER research in the Arabic language. The shared task encompasses subtask1 (FlatNER) and subtask2 (NestedNER). For this competition, we grant participants access to the Wojood corpus (Jarrar et al., 2022)[2], a substantial and diverse Arabic NER dataset known as Wojood. As shown in Figure 1, Wojood is particularly notable for its scale, containing approximately 550K tokens. About 12% of the corpus was collected from social media in *Pales-*

Mustafa Jarrar, Muhammad Abdul-Mageed, Mohammed Khalilia, Bashar Talafha, AbdelRahim Elmadany, Nagham Hamad, and Alaa Omer. 2023. WojoodNER 2023: The First Arabic Named Entity Recognition Shared Task. In Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023).

WojoodNER-2023 Shared Task

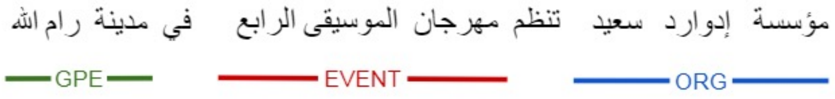
- WojoodNER-2023 is recognized as the inaugural shared task in Arabic Named Entity Recognition (NER).
- 88 % of Wojood contains nine different domains:
  - Health, finance, politics, ICT, terrorism, migration, history and culture, and law and elections.
- 12% of the corpus was collected from social media in Palestinian and Lebanese dialects.
- **550K tokens**



Topics in the Wojood NER corpus.

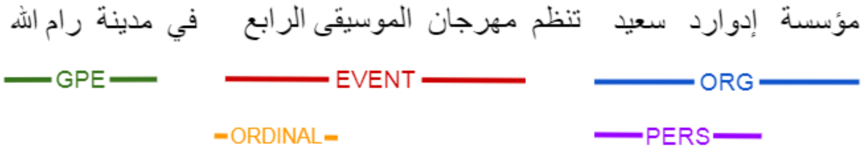
Task Description

- Subtask1 – FlatNER: each token in the data is labeled with only one tag. A flat NER dataset is derived from the nested NER.


مؤسسة إدوارد سعيد تنظم مهرجان الموسيقى الرابع في مدينة رام الله
ORG
EVENT
GPE

Flat NER example

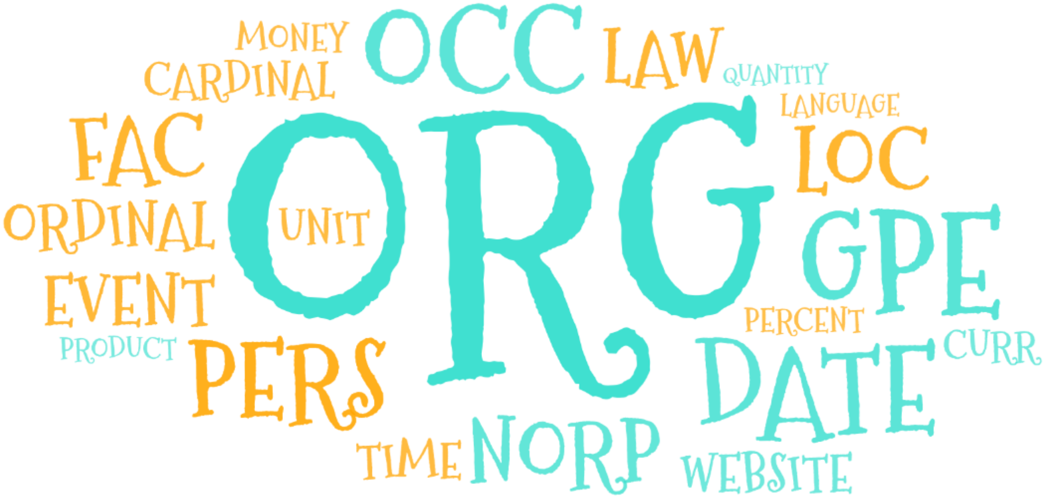
- Subtask2 – NestedNER: each token can have one or more tags


مؤسسة إدوارد سعيد تنظم مهرجان الموسيقى الرابع في مدينة رام الله
ORG
EVENT
GPE
PERS
ORDINAL

Nested NER example

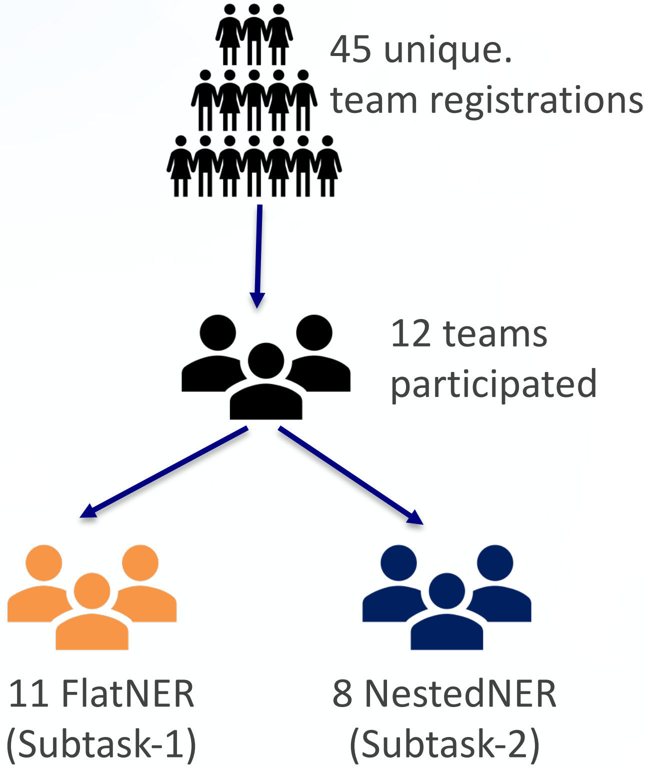
Shared Task Datasets

- **Splitting**: The data is split in to 70%, 10%, and 20% for training, development, and test dataset, respectively at the domain level.



21 Entity types in the Wojood dataset

Shared Task Teams & Results



45 unique. team registrations

12 teams participated

11 FlatNER (Subtask-1)

8 NestedNER (Subtask-2)

Team	Affiliation	Task
Alex-U 2023 NLP (Hussein et al., 2023)	Alexandria University	1,2
AlexU-AIC (Elkordi et al., 2023)	Alexandria University	1,2
AlphaBrains (Ehsan et al., 2023)	University of Gujrat, Pakistan	1,2
ARATAL	IPSA	1
El-Kawaref (Elkaref and Elkaref, 2023)	German University in Cairo	1
ELYADATA (Laouirine et al., 2023)	ELYADATA	1,2
Fraunhofer IAIS	Fraunhofer IAIS	1
LIPN (El Khbir et al., 2023)	LIPN, Université Paris 13	1,2
Lotus (Li et al., 2023)	MBZUAI	1,2
R00	Jordan University of Science and Technology	1,2
Think NER	Ulm University	1,2
UM6P & UL (El Mahdaouy et al., 2023)	Mohammed VI Polytechnic University	1,2

List of teams that participated in either one or both subtasks.

Shared Task Teams & Results

- All the models submitted to the shared task adopt the transfer learning approach, leveraging pre-trained models trained on various data sources.

- The top-performing models addressed the challenge of identifying nested entities of the same type

Rank	Team	F1	Precision	Recall
1	LIPN	91.96	92.56	91.36
2	El-Kawaref	91.95	91.43	92.48
3	ELYADATA	91.92	91.88	91.96
4	Alex-U 2023 NLP	91.80	91.61	92.00
5	Think NER	91.25	90.76	91.73
6	ARATAL	91.13	90.49	91.77
7	UM6P & UL	91.13	90.70	91.57
8	AlexU-AIC	91.13	91.33	90.92
	Baseline-I (ARBERTv2)	89.20	88.32	90.09
	Baseline-II (AraBERTv2)	87.33	86.00	88.00
9	AlphaBrains	87.15	87.45	87.58
10	Lotus	83.39	80.90	86.04
11	R00	76.99	76.67	77.31
12	Fraunhofer IAIS	64.45	65.53	63.40

Results of Subtask 1 - FlatNER

Shared Task Teams & Results

- All the models submitted to the shared task adopt the transfer learning approach, leveraging pre-trained models trained on various data sources.

- The top-performing models addressed the challenge of identifying nested entities of the same type

Results of Subtask 2 - NestedNER

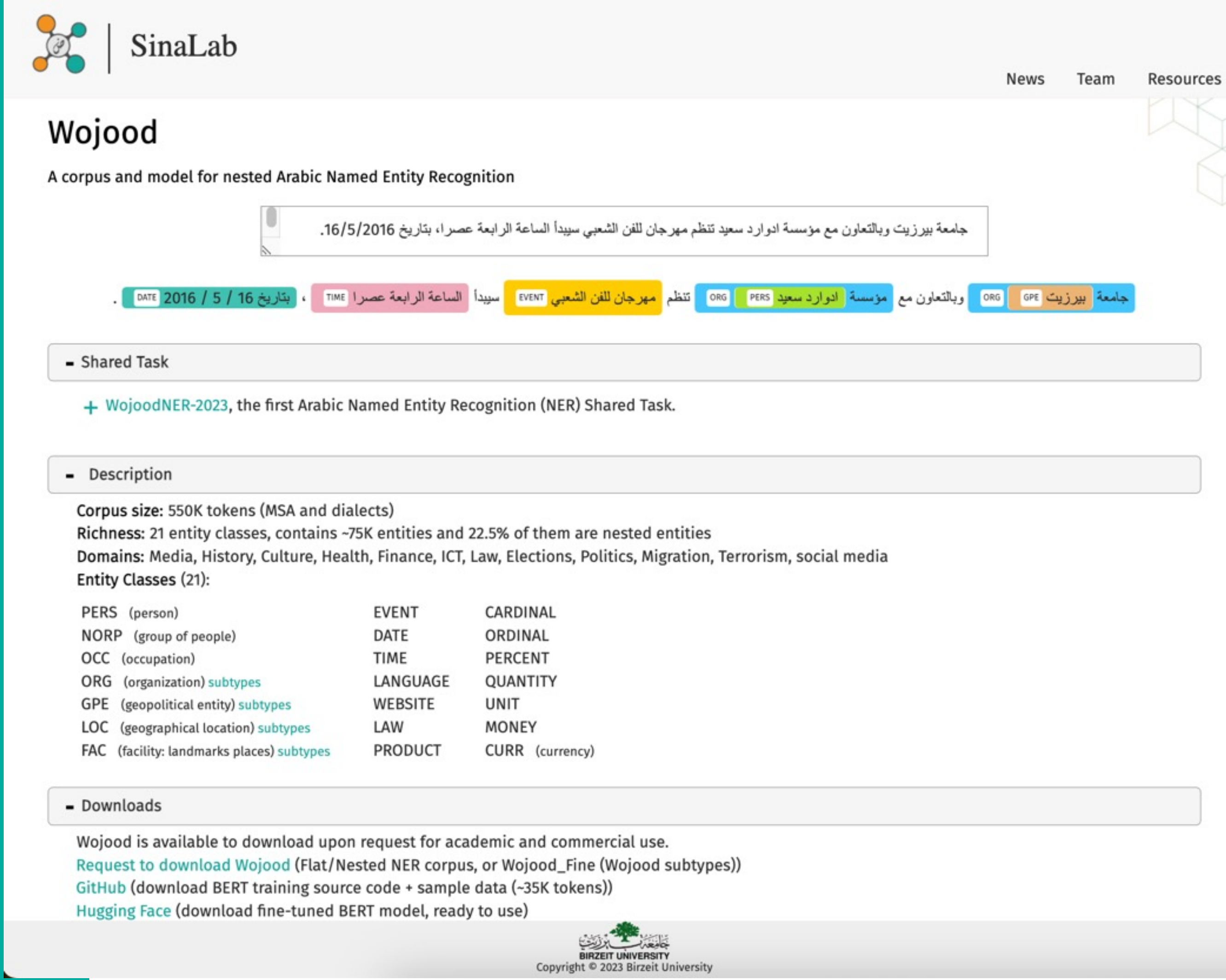
Rank	Team	F1	Precision	Recall
1	ELYADATA	93.73	93.99	93.48
2	UM6P & UL	93.03	92.46	93.61
3	AlexU-AIC	92.61	92.10	93.13
4	LIPN	92.45	92.31	92.59
	Baseline-I (ARBERTv2)	91.68	91.01	92.35
5	Think NER	91.4	90.03	92.82
	Baseline-II (AraBERTv2)	91.06	90.74	91.38
6	Alex-U 2023 NLP	90.01	89.39	90.63
7	AlphaBrains	88.84	88.45	89.23
8	Lotus	76.02	82.19	70.72

Post-Evaluation

Download Datasets

ArabicNER

ArabicNER-Wojood



Wojood

A corpus and model for nested Arabic Named Entity Recognition

جامعة بيرزيت وبالتعاون مع مؤسسة ادوارد سعيد تنظم مهرجان للفن الشعبي سيبدأ الساعة الرابعة عصرا، بتاريخ 16/5/2016.

جامعة بيرزيت GPE ORG وبالتعاون مع مؤسسة ادوارد سعيد PERS ORG تنظم مهرجان للفن الشعبي EVENT سيبدأ الساعة الرابعة عصرا TIME ، بتاريخ 16 / 5 / 2016 DATE .

- Shared Task

+ WojoodNER-2023, the first Arabic Named Entity Recognition (NER) Shared Task.

- Description

Corpus size: 550K tokens (MSA and dialects)
Richness: 21 entity classes, contains ~75K entities and 22.5% of them are nested entities
Domains: Media, History, Culture, Health, Finance, ICT, Law, Elections, Politics, Migration, Terrorism, social media
Entity Classes (21):

PERS (person)	EVENT	CARDINAL
NORP (group of people)	DATE	ORDINAL
OCC (occupation)	TIME	PERCENT
ORG (organization) subtypes	LANGUAGE	QUANTITY
GPE (geopolitical entity) subtypes	WEBSITE	UNIT
LOC (geographical location) subtypes	LAW	MONEY
FAC (facility: landmarks places) subtypes	PRODUCT	CURR (currency)

- Downloads

Wojood is available to download upon request for academic and commercial use.
Request to download Wojood (Flat/Nested NER corpus, or Wojood_Fine (Wojood subtypes))
GitHub (download BERT training source code + sample data (~35K tokens))
Hugging Face (download fine-tuned BERT model, ready to use)

BIRZEIT UNIVERSITY

https://ontology.birzeit.edu/

References

1. Mustafa Jarrar, Sanad Malaysha, Tymaa Hammouda, Mohammad Khalilia: 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.
2. Amal Nayouf, Tymaa Hammouda, Mustafa Jarrar, Fadi zaraket, Mohamad-Bassam Kurdy: Nâbra: Syrian Arabic Dialects with Morphological Annotations. In Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023. ACL.
3. Mustafa Jarrar, Ahmet Birim, Mohammed Khalilia, Mustafa Erden, and Sana Ghanem: ArBanking77: Intent Detection Neural Model and a New Dataset in Modern and Dialectical Arabic. In Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023. ACL.
4. Haneen Liqreina, Mustafa Jarrar, Mohammed Khalilia, Ahmed Oumar El-Shangiti, Muhammad AbdulMageed: Arabic Fine-Grained Entity Eecognition. In Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023. ACL.
5. Mustafa Jarrar, Muhammad Abdul-Mageed, Mohammed Khalilia, Bashar Talafha, AbdelRahim El-madany, Nagham Hamad, Alaa' Omar: WojoodNER 2023: The First Arabic Named Entity Recognition Shared Task. In Proceedings of the 1st Arabic Natural Language Processing Conference (Arabic- NLP), Part of the EMNLP 2023. ACL.
6. Nouran Khallaf, Elin Arfon, Mo El-Haj, Jon Morris, Dawn Knight, Paul Rayson, Tymaa Hammouda, Mustafa Jarrar: Open-source thesaurus development for under-resourced languages: a Welsh case study. The 4th LDK Conference on Language, Data and Knowledge, Vienna, Austria, 12-15 September 2023
7. Nagham Hamad, Mustafa Jarrar, Mohammad Khalilia, Nadim Nashif: Offensive Hebrew Corpus and Detection using BERT. The 20th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA). IEEE. Egypt. 2023
8. Sana Ghanem, Mustafa Jarrar, Radi Jarrar, Ibrahim Bounhas: A Benchmark and Scoring Algorithm for Enriching Arabic Synonyms. The 12th International Global Wordnet Conference (GWC2023), Global Wordnet Association. (pp.). San Sebastian, Spain, 2023
9. Sanad Malaysha, Mustafa Jarrar, Mohammad Khalilia: Context-Gloss Augmentation for Improving Arabic Target Sense Verification. The 12th International Global Wordnet Conference (GWC2023), Global Wordnet Association. (pp.). San Sebastian, Spain, 2023
10. Mustafa Jarrar, Mohammed Khalilia, Sana Ghanem: 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. 2022
11. Mustafa Jarrar, Fadi Zaraket, Tymaa Hammouda, Daanish Masood, Martin Waehlisch: Lisan: Yemeni, Iraqi, Libyan, and Sudanese Arabic Dialect Corpora with Morphological Annotations. The 20th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA). Pages(-). IEEE. Egypt. 2023 arXiv, DOI 10.48550/ARXIV.2212.06468. 2023
12. Karim El Haff, Mustafa Jarrar, Tymaa Hammouda, Fadi Zaraket: Curras + Baladi: Towards a Levantine Corpus. In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2022), Marseille, France. 2022
13. Mustafa Jarrar: The Arabic Ontology - An Arabic Wordnet with Ontologically Clean Content. Applied Ontology Journal, 16:1, 1-26. IOS Press. 2021
14. Moustafa Al-Hajj, Mustafa Jarrar: 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). PP 40--48, 2021
15. Moustafa Al-Hajj, Mustafa Jarrar: LU-BZU at SemEval-2021 Task 2: Word2Vec and Lemma2Vec performance in Arabic Word-in-Context disambiguation. In Proceedings of the Fifteenth Workshop on Semantic Evaluation (SemEval2021) Task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation (MCL-WiC). PP 748--755, Association for Computational Linguistics. 2021
16. Eman Naser-Karajah, Nabil Arman, Mustafa Jarrar: Current Trends and Approaches in Synonyms Extraction: Potential Adaptation to Arabic. In Proceedings of the 2021 International Conference on Information Technology (ICIT). PP 748--755, Association for Computational Linguistics. pp. 428-434, IEEE. 2021
17. Mustafa Jarrar, Eman Karajah, Muhammad Khalifa, Khaled Shaalan: Extracting Synonyms from Bilingual Dictionaries. The 11th International Global Wordnet Conference (GWC2021), Global Wordnet Association. (pp. 215-222). Pretoria, South Africa, 2021
18. 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, Hamdy Mubarak: A Panoramic Survey of Natural Language Processing in the Arab World. Communications of the ACM, April 2021, Vol. 64 No. 4, Pages 72-81
19. Mustafa Jarrar: Digitization of Arabic Lexicons. Arabic Language Status Report. UAE Ministry of Culture and Youth. Pages 214-2017. Dec 2020
20. Mustafa Jarrar, Hamzeh Amayreh: An Arabic-Multilingual Database with a Lexicographic Search Engine. The 24th International Conference on Applications of Natural Language to Information Systems (NLDB 2019). Pages(234-246). LNCS 11608, Springer. 2019
21. Mustafa Jarrar, Hamzeh Amayreh, John P. McCrae: Representing Arabic Lexicons in Lemon - a Preliminary Study. The 2nd Conference on Language, Data and Knowledge (LDK 2019). Pages(29-33). CEUR, Volume 2402. ISSN:1613-0073. Leipzig, Germany. 2019
22. Diana Alhafi, Anton Deik, Mustafa Jarrar: Usability Evaluation of Lexicographic e-Services. The 16th IEEE/ACS International Conference on Computer Systems and Applications (AICCSA). Pages(1-7). IEEE. Abu Dhabi, UAE. 2019
23. Mustafa Jarrar, Fadi Zaraket, Rami Asia, Hamzeh Amayreh: Diacritic-Based Matching of Arabic Words. ACM Asian and Low-Resource Language Information Processing. Volume 18, No 2, Pages(10:1-10:21), ACM, ISSN:2375-4699. December, 2018
24. Mustafa Jarrar: Search Engine for Arabic Lexicons. The 5th Conference on Translation and the Problematics of Cross-cultural Understanding. The Forum for Arab and International Relations. Doha, Qatar. December, 2018
25. Diab Abuaiadah, Dileep Rajendran, Mustafa Jarrar: Clustering Arabic Tweets for Sentiment Analysis. The 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications. Pages(499-506). IEEE Computer Society. ISBN:9781538635810. (doi.10.1109/AICCSA.2017.162). Hammamet, Tunisia. 2017
26. Mustafa Jarrar, Nizar Habash, Faeq Alrimawi, Diyam Akra, Nasser Zalmout: Curras: An Annotated Corpus for the Palestinian Arabic Dialect. Journal Language Resources and Evaluation. Pages(745-775). Volume(51)