The 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023. ACL.
Singapore, 7 December, 2023

Arabic Fine-Grained Entity Recognition

Haneen Liqreina **Mustafa Jarrar** **Mohammad Khalilia**

Birzeit University
Palestine

Ahmed El-Shangiti **Muhammad Abdul-Mageed**

UBC and MBZUAI
Canada

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

ties, 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,

Haneen Abdallatif Liqreina, Mustafa Jarrar, Mohammed Khalilia, Ahmed Oumar El-Shangiti, Muhammad Abdul-Mageed. Arabic Fine-Grained Entity Recognition. In The First Arabic Natural Language Processing Conference, 2023.

http://www.jarrar.info/publications/LJKOA23.pdf

Natural Language Understanding Tools and Datasets

Open Source

https://sina.birzeit.edu/resources

SinaLab

News Team 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

Contributions

$\text{Wojood}_{\text{Fine}}$

We extended the Wojood NER corpus by adding subtypes of entities for each of {ORG, GPE, LOC,FAC}

$\text{Wojood}_{\text{Fine}}$
Fine-Grain and Nested Arabic Named Entities Corpus

نيامي هي عاصمة نيجيريا

Niamey is the capital of Nigeria

GPE ▸ Country

لتتوافق وأولويات نيجيريا في المنطقة

To align with the Nigeria's priorities in the region

GPE ▸ GPE_ORG

$\text{Wojood}_{\text{Fine}}$
Fine-Grain and Nested Arabic Named Entities Corpus

صورة لبعض المرضى في مستشفى الشفاء

A picture of some patients at Al-Shifa Hospital

ORG ▸ ORG_FAC

Wojood$_{Fine}$

Fine-Grain and Nested Arabic Named Entities Corpus

أعيش في شمال شرق مدينة غزة

I live northeast of Gaza City

GPE ▸ Town

LOC ▸ Region-General

$\text{Wojood}_{\text{Fine}}$
Fine-Grain and Nested Arabic Named Entities Corpus

جامعة بيرزيت تعلن عن مجموعة من المنح لطلبة الدكتوراة

Birzeit University announces a set of scholarships for PhD students

GPE ▸ Town

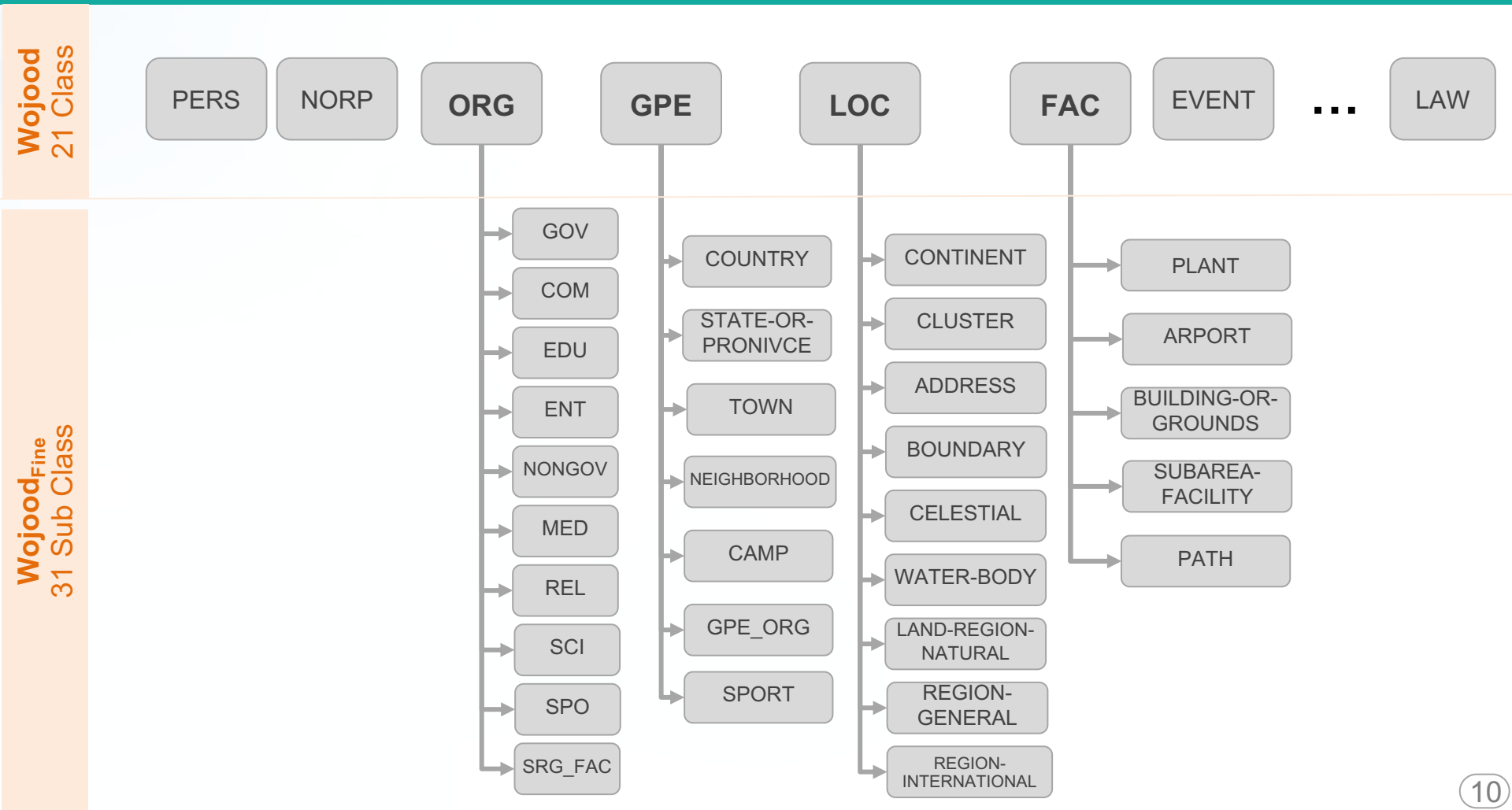
ORG ▸ EDU

Wojood Vs Wojood_Fine Counts

Before adding subtypes, we had to revise the annotations of {GPE,ORG,LOC,FAC} with new guidelines.

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**

Subtypes


Wojood
21 Class
PERS
NORP
ORG
GPE
LOC
FAC
EVENT
...
LAW
Wojood Fine
31 Sub Class
GOV
COM
EDU
ENT
NONGOV
MED
REL
SCI
SPO
SRG_FAC
COUNTRY
STATE-OR-PRONIVCE
TOWN
NEIGHBORHOOD
CAMP
GPE_ORG
SPORT
CONTINENT
CLUSTER
ADDRESS
BOUNDARY
CELESTIAL
WATER-BODY
LAND-REGION-NATURAL
REGION-GENERAL
REGION-INTERNATIONAL
PLANT
ARPORT
BUILDING-OR-GROUNDS
SUBAREA-FACILITY
PATH

Annotation Guidelines

10 sub-entity classes for **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).mentions of the entireties of any of the seven continents.	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.

Annotation Guidelines

7 sub-entity classes for **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.

محافظة القاهرة، قطاع غزة، إقليم كردستان، لواء نابلس.

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.

أصدرت الولايات المتحدة تقريرها، قررت فلسطين إعفاء المتضررين.

SPORT

Athletes, Sports Teams.

}مباراة المغرب، الفرق الرياضية، برشلونة، ميلان.

Annotation Guidelines

9 sub-entity classes for **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"). 17، 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.

Annotation Guidelines

5 sub-entity classes for **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. جزء من مبنى، غرفة ،زنزانة.
SUBAREA-FACILITY	Taggable portions of facilities. منزل، مبنى، مستشفى، معبر.
PATH	Streets, canals, and bridges. الشوارع الرئيسية، الخطوط الهاتفية، الحواجز.

NER Results

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	$0.883^{\pm 0.00}$	$\mathbf{0.885}^{\pm 0.00}$

M1: ARBERTv2 M2: MARBERTv2 M3: ARABERTv2

Results of fine-tuned models on the three different tasks. The results are represented as F1 averaged over 3 runs.

Limitations and Out-of-Domain Performance

- Unseen domains and different time periods
- compiled from Aljazeera news articles published in 2023
- Each domain about 3k tokens

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**$^{\pm 0.08}$

M1: ARBERTv2 M2: MARBERTv2 M3: ARABERTv2

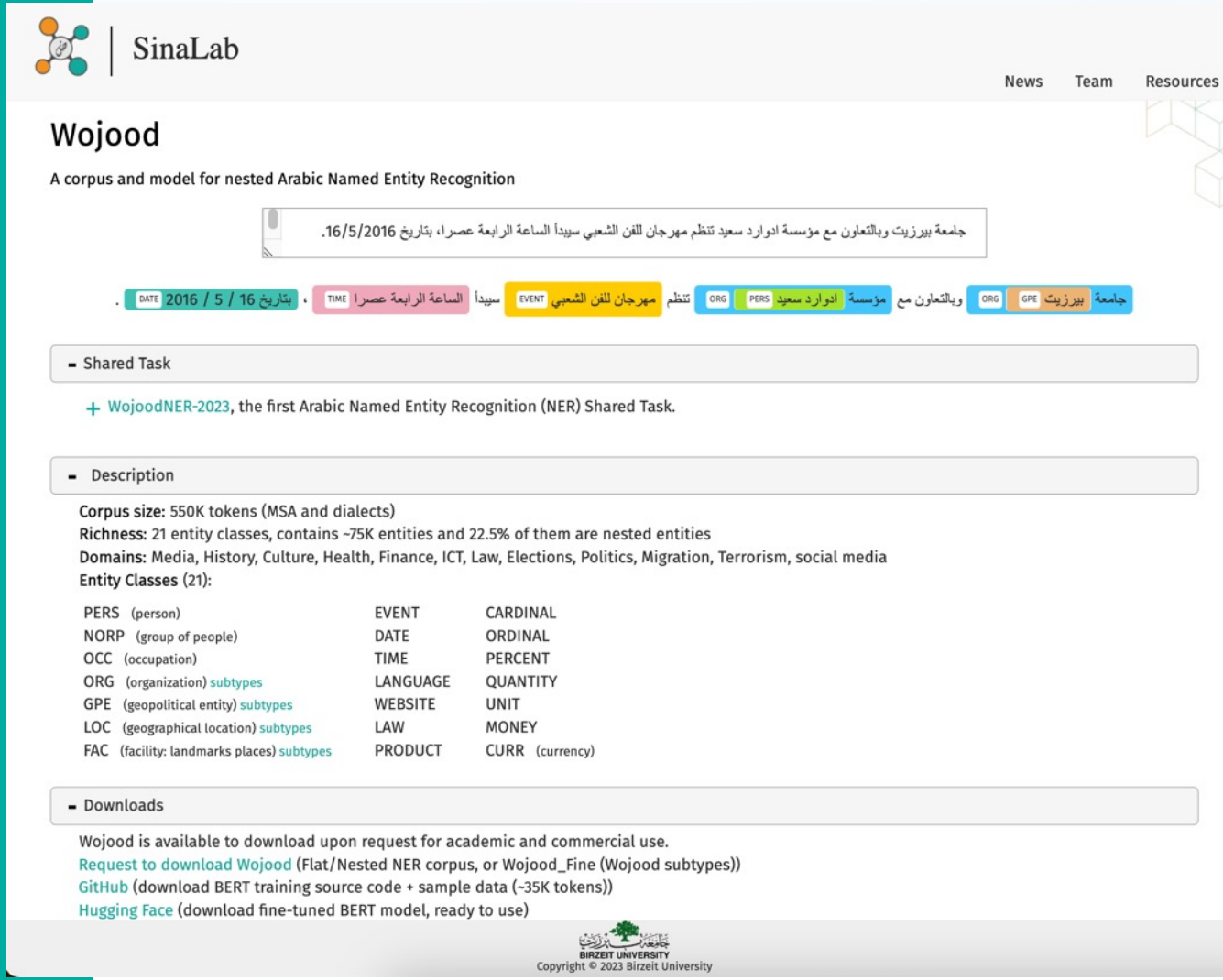
The results are represented as F1 averaged over 3 runs.

Try

ArabicNER

ArabicNER-Wojood

https://ontology.birzeit.edu/wojood



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).