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

SALMA: Arabic Sense-Annotated Corpus and WSD Benchmarks




Mustafa Jarrar **Sanad Malaysha** **Tymaa Hammoudah** **Mohammed Khalilia**
SinaLab, Birzeit University
Palestine

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

Semantic Understanding Tasks

1. **WSD** Word Sense-Disambiguation

2. **TSV** Target Sense Verification

3. **WiC** Word-in-Context

historically known to be the most diffecult NLP tasks

WSD - Word-Sense Disambiguation

Given a word in a context and a set of sense for this word, which sense this word denotes?

قصيدة من عيون الشعر

Set of senses

1. **عُضو الإبصار** في الإنسان والحيوان: له عينان كعَيْنَيْ الصقر - ألا إنّما العينان للقلب رائدٌ ...
2. **جاسوس**، "كان عينًا لدولةٍ أجنبيَّة . بثَّ العيون : تجسَّس، راقب - فلانٌ عَيْن على فلان : ناظر عليه
3. **أجود كلّ شيء** وأحسنه ونفيسُه: عيونُ الفنِّ.
4. **حارس:** فلان عين على المكان.
5. **الحاضر من كل شيء** أصبح أثرًا بعد عين ...
6. **عَيْنُ الماء**:- ينبوعه، تُحلِّق الطّيورُ فوق عيون الماء
7. **عَيْن الشَّيء**:- نفسه، ذاته (تستعمل للتوكيد): جاء القوم أعينهم - كنّا في المكان عينه.
8. **عَيْن العقل**:- قدرة ذهنيَّة مَوْروثة على التخيُّل وتذكّر الأحداث.
9. .....

TSV - Target Sense Verification

- Given a context, target word and gloss, TSV aims to decide whether it is true that this gloss is the intended meaning of the target in this context.
- Whether a (Context-Gloss pair) is true or false

Example:

Context	Gloss	Label
تمشي بين الجداول والازهار Walking among streams and flowers	مجرى صغير متفرع من نهر A small stream branching from a river	True
تمشي بين الجداول والازهار Walking among streams and flowers	تنظيم للبيانات والمعلومات في صورة صفوف وأعمدة Organization of data in the form of rows and columns	False

WiC - Word-in-Context

Determines whether a target word in two contexts (sentences) is used in the same sense or not

Example:

Context 1	Context 2	Label
تمشي بين الجداول والازهار Walking among streams and flowers	كنا نمرح ونستمتع بجداول الربيع We were playing and enjoying the spring streams	True
تمشي بين الجداول والازهار Walking among streams and flowers	انظر الجداول في الصفحة الثالثة See the table in third page	False

SALMA: Arabic Sense-Annotated Corpus and WSD Benchmarks

Mustafa Jarrar, Sanad Malaysha, Tymaa Hammouda, Mohammed Khalilia
Birzeit University, Palestine
{mjarrar, smalaysha, thammouda, mkhalilia}@birzeit.edu

Abstract

SALMA, the first Arabic sense-annotated corpus, consists of ~34K tokens, which are all sense-annotated. The corpus is annotated using two different sense inventories simultaneously (Modern and Ghani). SALMA novelty lies in how tokens and senses are associated. Instead of linking a token to only one intended sense, SALMA links a token to multiple senses and provides a score to each sense. A smart web-based annotation tool was developed to support scoring multiple senses against a given word. In addition to sense annotations, we also annotated the corpus using six types

1949/1955), but it has recently gained more attention due to the advances in learning contextualized word representations from language models, such as BERT (Devlin et al., 2019) and GPT (Radford et al., 2018).

As glosses are short descriptions of senses (Jarrar, 2006, 2005), recent research has demonstrated promising results in WSD task by framing the problem as a sentence-pair (context-gloss) binary classification task, referred to as Target Sense Verification (TSV), where the context is a sentence con-

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.

http://www.jarrar.info/publications/JMHK23.pdf

SALMA
Arabic Sense-Annotated Corpus

❖ SALMA Corpus: First Arabic sense-annotated corpus (~34K tokens)

- Annotated using **two sense inventories** (Modern and Ghani)
- Annotated using **six types of named entities**

Corpus	Unique Senses	Annotation Type	Corpus Size (tokens)	Annotations				
				Nouns	Verbs	Func. Words	Punc.+ Digits	Total
SALMA (ours)	4,151 word senses (from each sense inventory) 6 types of named entities	all senses of all words	34K	**19,030**	**2,763**	**7,116**	**5,344**	**34,253**

❖ SALMA System: First Arabic end-to-end WSD system

- Using TSV

❖ SALMA Baseline: Compute the WSD baseline in different settings

- Baseline = 84.2%

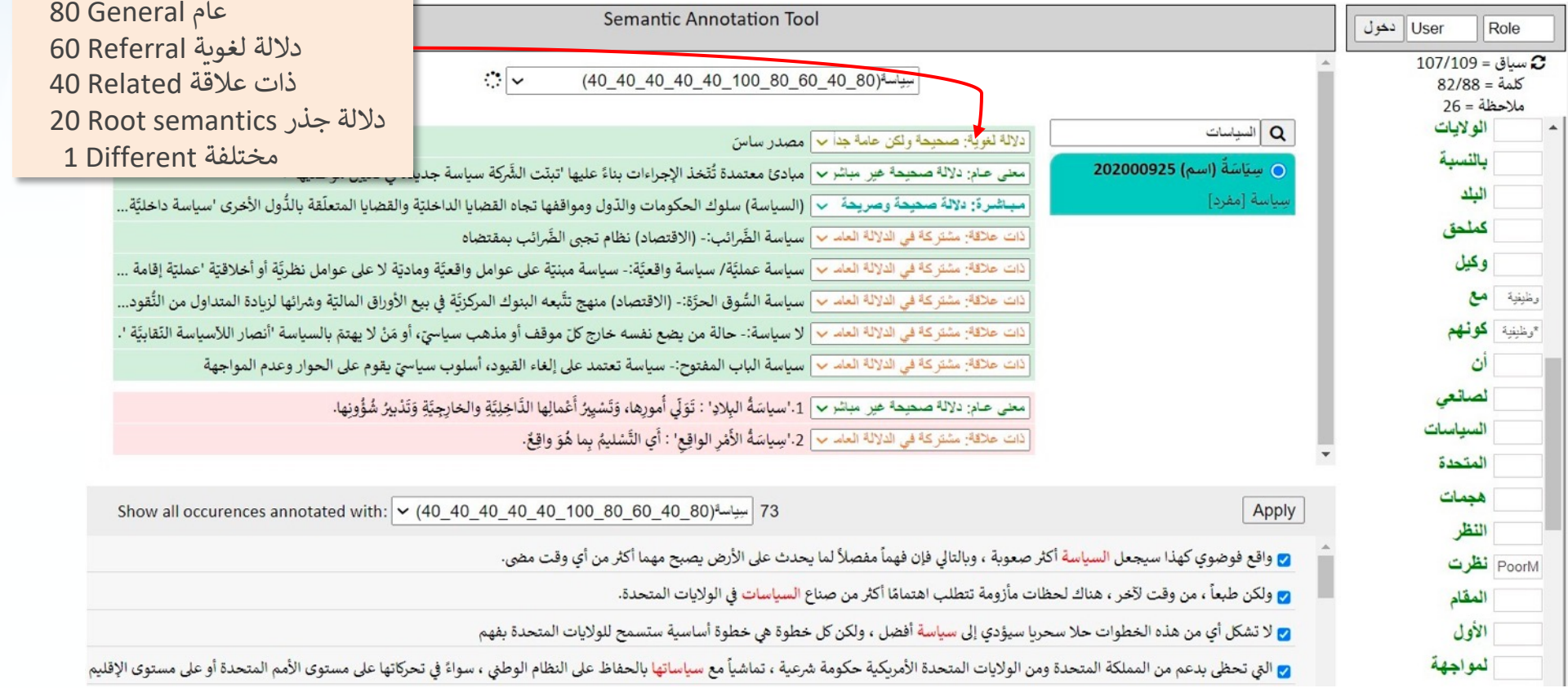
Corpus Collection

- SALMA corpus (34K token) is part of Wojood corpus (
- Collected from 33 online media sources written in MSA and covering general topics

Annotation Tool & Guidelines

Guidelines

100 Explicate مباشرة
80 General عام
60 Referral دلالة لغوية
40 Related ذات علاقة
20 Root semantics دلالة جذر
1 Different مختلفة



Statistics

SALMA

Term	Noun	Verb	Func. Words	Punc+ Digits	Total
Tokens	19,030	2,763	7,116	5,344	**34,253**
Unique Tokens	6,670	1,593	322	175	**8,760**
Unique Lemmas	2,904	677	119	175	**3,875**
Unique Senses	3,151	792	206	2	**4,151**

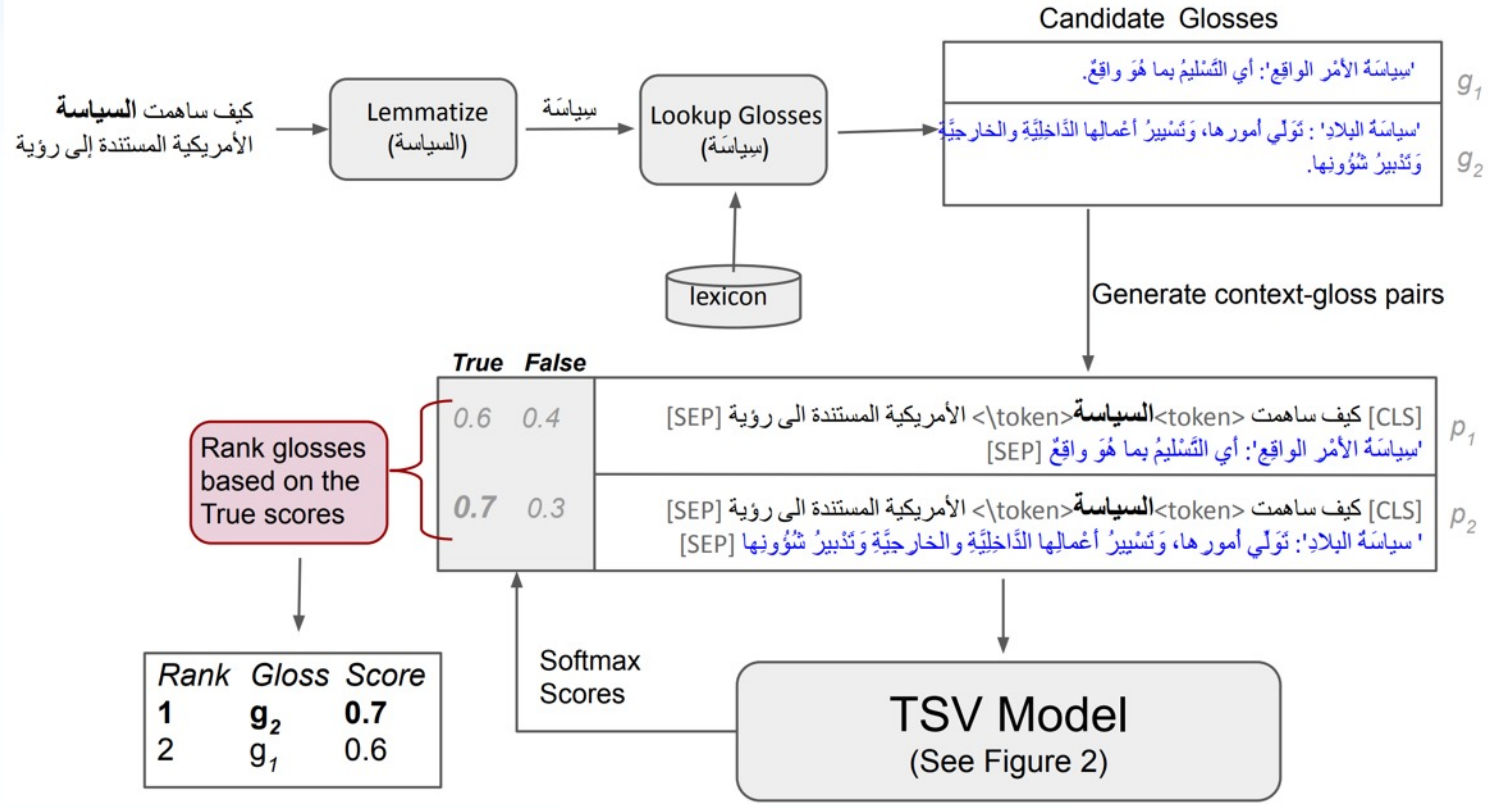
Named Entities

Tag	Named Entity Mentions	Tokens in the Entity Mentions
PERS	294	568
ORG	1,123	2,108
GPE	1,086	1,295
LOC	166	318
FAC	22	59
CURR	37	41
Total	**2,728**	**4,389**

Coverage

Term	Modern	Ghani
Lemmas	80% (2,788/3,522)	78% (2,724/3,522)
Senses (Without Proper nouns)	83% (3,430/4,151)	78% (3,226/4,151)
Proper Nouns Senses	4% (9/213)	14% (30/213)

SALMA
End-to-end WSD system (using TSV)



WSD Baselines

Different TSV models

TSV Model	Lexicons	Accuracy
Razzaz	Modern Ghani	66.0% 68.4%
ArabGlossBERT	Modern Ghani	**84.2%** 77.6%
Aug-ArabGlossBERT(D9)	Modern Ghani	82.6% 78.7%

Focus on ArabGlossBERT TSV model

Window	Lexicon	Accuracy Target Sense Rank			Accuracy (Top1) per POS		
		Top1	Top2	Top3	Noun	Verb	Func.
All	Modern	82.8	94.2	97.4	83.5	77.9	41.2
	Ghani	77.0	89.3	94.1	78.5	66.0	36.0
11	**Modern**	**84.2**	95.1	98.1	85.4	76.1	37.9
	Ghani	**77.6**	90.1	94.9	79.4	61.7	31.8
9	Modern	83.5	95.0	97.9	84.4	78.3	37.7
	GHani	77.3	90.1	94.8	79	63.7	32.2
7	Modern	83.8	95.1	97.9	84.8	77.4	38.9
	Ghani	77.3	90.0	94.9	79.1	62.9	31.8
5	Modern	84.0	95.1	98.1	85.3	75.6	40.0
	Ghani	77.6	90.1	94.9	79.5	61.6	31.7
3	Modern	82.8	94.4	97.6	84.4	71.8	42.1
	Ghani	77.4	90.0	94.8	79.4	59.7	32.1

Try

https://sina.birzeit.edu/salma/

SinaLab

News Team Resources

SALMA سلمى

A corpus and model for Arabic Word Sense Disambiguation (WSD).
Version: 1.0 (updated on 22/10/2023)

قصيدة من عيون الشعر

WSD

◂ **قصيدة** (قَصِيدَة 1_303044571): مجموعة من الأبيات الشِّعريّة متّحدة في الوزن والقافية والرَّويّ وهي تتكوّن من سبعة أبيات فأكْثر "قصيدة غزليّة ".

بَيْتُ القصيدة : البيت المتضمّن غاية الشّاعر، أو أنفس أبياتها، أو مثل يُضرب في تفضيل بعض الشَّيء على كُلّه - مطلع القصيدة : أوّل بيت منها

◂ **من** (مِن 1)

◂ **عيون** (عَيْن 2_303038477): أجود كلّ شيء وأحسنه ونفيسُه "قصيدة من عيون الشّعر - عيونُ الفنِّ ".

◂ **الشعر** (شِعْر 1_303029103): كلام موزون مقفّى قصدًا يعتمد على التخييل والتأثير؛ ليوحي بإحساسات مؤثِّرة وصور خياليّة "شعر صافي الديباجة - نظم الشِّعْرَ - ما الشِّعْر إلاّ شعورُ المرء يُرسلُه ... عفو البديهة عن صدق وإيمانِ - إنَّ مِنَ الشِّعْرِ لَحِكْمَةً [حديث] - حوَمَا عَلَّمْنَاهُ الشِّعْرَ وَمَا يَنْبَغِي لَهُ> يس/ 69 ". أنشده الشِّعرَ : قرأه عليه - أوابِدُ الشِّعْر : ما لا تُماثَل جودتُه أو قوافيه الشاردة - ربَّة الشِّعْر : إلهة الشِّعْر عند الوثنيِّين - شطرا بيت الشِّعْر : الصدر

+ Description

+ Downloads

BIRZEIT UNIVERSITY

Try

https://sina.birzeit.edu/salma/

SinaLab

News Team Resources

SALMA سلمى

A corpus and model for Arabic Word Sense Disambiguation (WSD).
Version: 1.0 (updated on 22/10/2023)

صورة لعيون جميله

WSD

◂ **صورة** (صُورَة 1 303032440): (الطبيعة والفيزياء) ما تراه العين مباشرة أو من خلال عدسة أو في مرآة أو مرتدّ عنها على سطح ما

◂ **لعيون** (عَيْن 2 303038475): (التشريح) عُضو الإبصار في الإنسان والحيوان "له عينان كعَيْنَيْ الصقر - ألا إنّما العينان للقلب رائدٌ ... فما تألفُ العينان فالقلب آلِفُ - حفَرَجَعْنَاكَ إِلَى أُمِّكَ كَيْ تَقَرَّ عَيْنُهَا> طه/ 40 - حوَلِتُصْنَعَ عَلَى عَيْنِي> طه/ 39 : لتصنع تحت رعايتي وحفظي وإكرامي". أخَذ بعَيْن الاعتبار : قدّر، راعى أمرًا ما - أصابته العَيْنُ : حُسِد - أغمض عَيْنه عنه : تجاهله، تغافله - أنت على عَيْني : يقال في الإكرام والحفظ جميعًا - إنسان العَيْن :

+ Description

+ Downloads

BIRZEIT UNIVERSITY

Try

https://sina.birzeit.edu/salma/

SinaLab

News Team Resources

SALMA سلمى

A corpus and model for Arabic Word Sense Disambiguation (WSD).
Version: 1.0 (updated on 22/10/2023)

اعمل بجامعة بيرزيت وأحب زيت الزيتون

WSD

◂ **اعمل** (عَمِلَ_1 303037736): عمِلَ الرَّجلُ:- مارس نشاطًا وقام بجهد للوصول إلى نتيجة نافعة "عمل بنظام - عمل على إرضاء والده - عمل للصالح العامّ ".
يُعمل بالقانون : يطبَّق ويُنفَّذ - يعمل عن بُعد : يمارس العمل عن طريق حاسوب في بيته متّصل بمكان عمله

◂ **بجامعة بيرزيت** اسم مؤسسة

◂ **وأحب** (أَحَبَّ_1 303009171): أحبَّ الشَّيءَ أو الشَّخصَ:- أحبَّه، ودّه ومال إليه، عكس كرهه "جئتك بقوم يحبُّون الموتَ كما تحبُّون الحياةَ يرغبون فيه ولا يخافونه - من أحبّ شيئًا أكثر من ذِكْره - سجِّل نصيحة من يحبّك وإن كنت لا تتقبّلها في حينها [مثل أجنبيّ] يماثله في المعنى المثل العربيّصديقك من صَدَقَك لا من صدّقَك - إنّ المُحِبَّ إذا أحبَّ حبيبَه ... صدَق الصَّفاءَ وأنجز الموعودا أخلص وصدق في مودّته - لاَ يُؤْمِنُ أَحَدُكُمْ حَتَّى يُحِبَّ لأَخِيهِ مَا يُحِبُّ لِنَفْسِهِ [حديث] يتمنّى - حقُلْ إِنْ كُنْتُمْ تُحِبُّونَ اللهَ فَاتَّبِعُونِي يُحْبِبْكُمُ اللهُ< آل عمران/ 31 ". المحبّ المخلص : الصّادق المحبَّة - كما تحبّ : حَسْب ما تريد أو ترغب

◂ **زيت الزيتون** (زَيتُ الزَّيتون 332001242): زيت ثابت يُستحصل من عصر ثمار الزيتون الناضجة. يُستخدم في التغذية وفي الصناعات الدوائية.

BIRZEIT UNIVERSITY

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).