Event-Arguments Extraction Corpus and Modeling using BERT for Arabic

Alaa Aljabari   Lina Duaibes   Mustafa Jarrar   Mohammed Khalilia

SinaLab, Birzeit University
Palestine

1

Open Source

**Arabic NLP
Tools and Datasets**

https://sina.birzeit.edu/resources

Resources

Download and demo our tools and datasests

+ Arabic Ontology	الأنطولوجيا العربية
+ Lexicographic Databases (Qabas)	حوسبة المعاجم (قبس و150 معجم)
+ Dialect Corpora (Currasat)	مدونة اللهجات العامية (كراسات)
+ SinaTools	أدوات سينا
+ Morphology Tagger (Alma)	المحلل الصرفي (ألمى)
+ Word Sense Disambiguation (Salma)	المحلل الدلالي (سلمى)
+ Named Entity Recognition (Wojood)	استخراج أسماء الاعلام (وجود)
+ Relation Extraction	استخراج العلاقات
+ Social Computing (Fada)	الإنسانيات الحاسوبية والتواصل الاجتماعي (فضا)
+ Synonyms	استخراج المترادفات
+ Chatbots and intent detection (AraBanking77)	المساعدات الآلية

Motivation

❖ Understanding and **extracting events is important in** NLP applications like **disaster monitoring** , **emergency response**, .. etc.

❖ A notable **gap in the availability of annotated corpora** for this purpose, especially for under-resourced languages like Arabic.

Event Argument Extraction

❖ **Input :** a sentence with predicted named entities.

❖ **Goal** : Extract relations between events and other entities (event arguments).
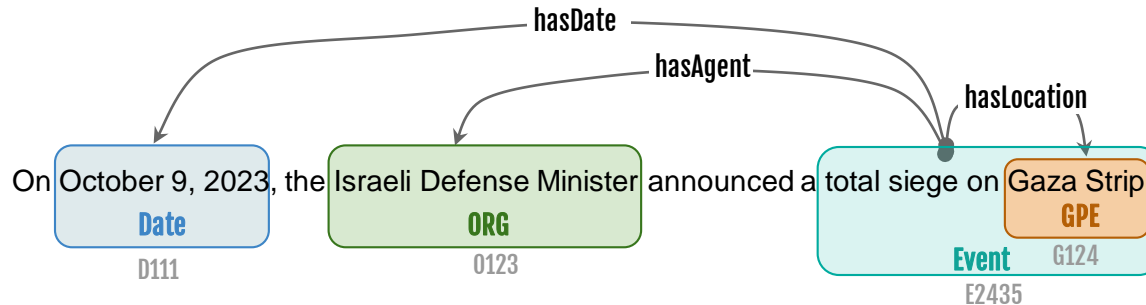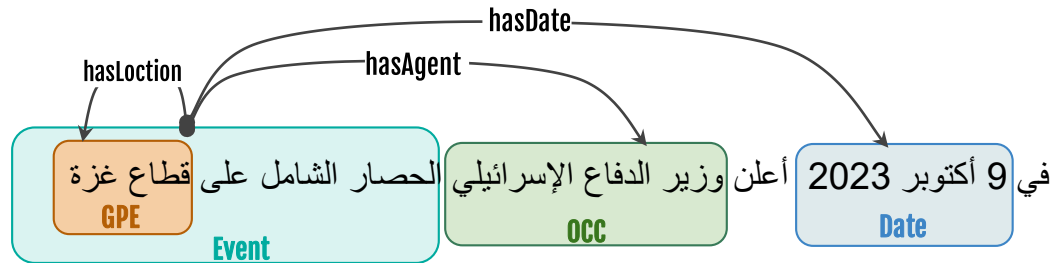
Our contribution

❖ $Wojood^{Hadath}$ corpus: A manually annotated corpus with 550k tokens for event argument relations, used to create $Hadath^{NLI}$ (25k premise-hypotheses pairs).

❖ **Novel Methodology for event-argument extraction** by framing it as an NLI problem, achieving high performance.

❖ $Wojood^{OutOfDomain}$: Additional annotated corpus (80k tokens) for **out-of-domain evaluation**.

❖ **End-to-End System** for **event-argument relation extraction**.

Wojood^{Hadath} corpus

Wojood^{Hadath} corpus

❖ Wojood is a rich Arabic nested named entity corpus with **550k tokens**. (Jarrar et al., 2022b)

❖ Contains **2,772 annotated events**.

❖ Extended the existing Wojood corpus to construct an event-argument corpus.

❖ Objective: Identify event arguments and establish relationships between these arguments and respective event entities.

Relationship Types

We annotated Wojood with the following relations:

❖ **hasAgent**: specifies participant(s) involved in the event, which can be a PERS, ORG,OCC, or NORP named entities.

❖ **hasLocation**: indicates where the event occurred, which can be GPE, LOC, and FAC named entities.

❖ **hasDate**: points when the event occurred, which can be TIME or DATE.

Corpus Statistics

❖ **Wojood**^{Hadath} comprise 1,974 annotated events with event-argument relations.

Relation	Count
hasAgent	423
hasLocation	833
hasDate	1332
Total	**2588**

Annotation Guidelines

❖ Event arguments are recognized only **within the same sentence**.

❖ **Different IDs for distinct entities**, even if they refer to the same person
[e.g. The killing of the Egyptian president Anwar al-Sadat]

❖ **An event can have multiple agents.**
[e.g. Signing a cooperation agreement between the Lebanese government and the Central Bank]

❖ Two event entities can **share the same argument** in a sentence.
[e.g. The political situation in Egypt is tense after the 1967 War (Al Naksa)] - sharing "the year 1967"

Inter-annotator Agreement

❖ **Randomly selecting 5%** of annotations for evaluation.

❖ Inter-annotator Agreement (IAA) using **Cohen's kappa** and **F1-score**.

Relation	TP	FN	FP	K	F1-Score
hasAgent	37	10	10	67.85%	79%
hasLocation	29	2	2	91.70%	94%
hasDate	43	2	6	87.15%	91%
Overall	**109 count**	**14 count**	**18 count**	**82.23% macro**	**87.20% micro**

Our contribution

❖ **Wojood**$^{\text{Hadath}}$ corpus: A manually annotated corpus with 550k tokens for event argument relations, used to create **Hadath**$^{\text{NLI}}$ (25k premise-hypotheses pairs).

Novel Methodology **for event-argument extraction** by framing it as an NLI problem, achieving high performance.

❖ **Wojood**$^{\text{OutOfDomain}}$: Additional annotated corpus (80k tokens) for **out-of-domain evaluation**.

❖ **End-to-End System** for **event-argument relation extraction**.

Novel Methodology for Event-Argument Extraction

Event Argument Extraction Modeling

Our proposal

Frame the **relation extraction** problem as **Natural Language Inference (NLI) task**
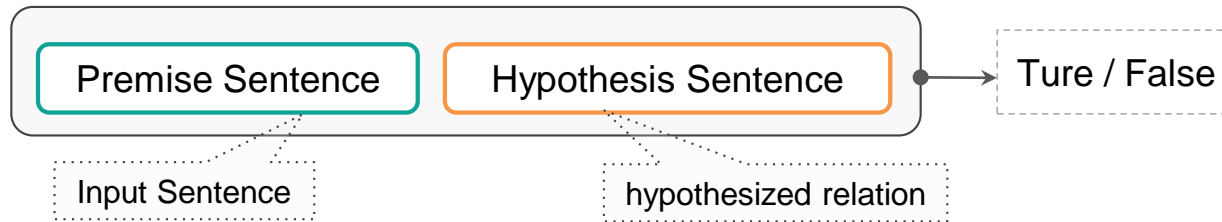
Event Argument Extraction Modeling

❖ Examples:

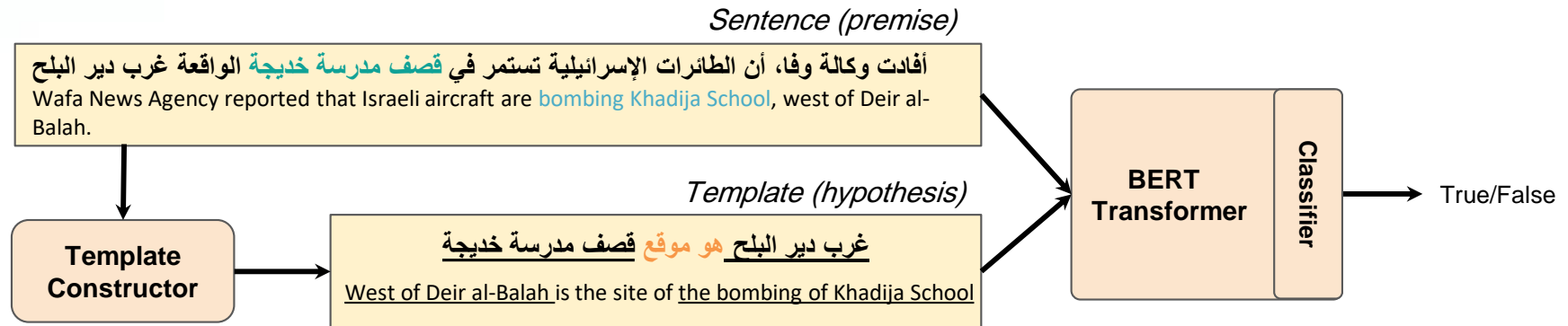| Premise	Hypothesis	Label
أفادت وكالة وفا، أن الطائرات الإسرائيلية تستمر في قصف مدرسة خديجة الواقعة غرب دير البلح Wafa News Agency reported that Israeli aircraft are bombing Khadija School, west of Deir al-Balah.	غرب دير البلح هو موقع قصف مدرسة خديجة West of Deir al-Balah is the site of the bombing of Khadija School	True
أفادت وكالة وفا، أن الطائرات الإسرائيلية تستمر في قصف مدرسة خديجة الواقعة غرب دير البلح Wafa News Agency reported that Israeli aircraft are continuing to bomb Khadija School, west of Deir al-Balah.	وكالة وفا أحد الفاعلين في قصف مدرسة خديجة Wafa Agency is one of the agents of the bombing of Khadija School	False

Event Argument Extraction Modeling

❖ Framing EAE as a Natural Language Inference (NLI) task.

Sentence (premise)

أفادت وكالة وفا، أن الطائرات الإسرائيلية تستمر في **قصف مدرسة خديجة** الواقعة غرب دير البلح
Wafa News Agency reported that Israeli aircraft are bombing Khadija School, west of Deir al-Balah.

Template (hypothesis)

غرب دير البلح **هو موقع** قصف مدرسة خديجة

West of Deir al-Balah is the site of the bombing of Khadija School

Template Constructor

BERT Transformer

Classifier

True/False

NLI Dataset

❖ Construct **Hadath**^NLI from **Wojood**^Hadath

➜ **Template Selection:** Based on entity type.

➜ **Positive Pairs:** Entities annotated as arguments of a specific event are used to generate a positive hypothesis.

➜ **Negative Pairs:** Entities that are not linked with events (i.e., not event arguments) are used to generate negative pairs.
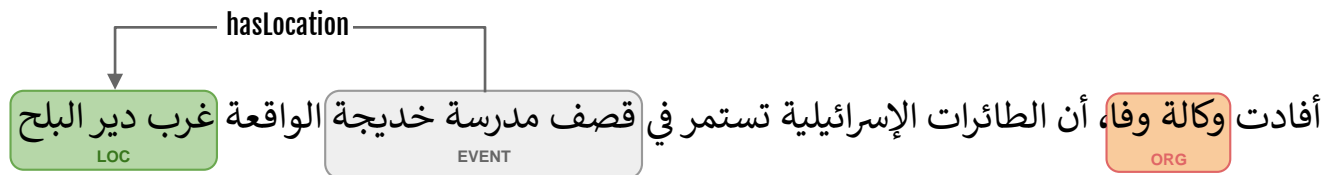
NLI Dataset

❖ Hadath[NLI] Statistics:

Phase	Pairs	Positive	Negative	Total
Train	hasAgent	1,248	6,156	7,404
	hasLocation	2,268	4,456	6,724
	hasDate	3,716	2,948	6,664
	SubTotal	**7,232**	**13,560**	**20,792**
Test	hasAgent	111	653	764
	hasLocation	267	464	728
	hasDate	403	318	72
	SubTotal	**778**	**1,435**	**2,213**
	Total	**8010**	**14,995**	**23,005**

Model Training

❖ Trained on **ArBERTv2***.*

❖ **Weighted** Cross Entropy and **Contrastive** Loss.

Results

❖ EAE Results on $\text{Hadath}^{\text{NLI}}$.

Class	Support	Precision	Recall	F1 Score
Positive	778	90.06%	92.42%	92.24%
Negative	1,435	95.99%	95.68%	95.78%
Average				**94.01%**

Promising results!!!

To validate generalization:
Out-of-Domain Evaluation

Our contribution

❖ $Wojood^{Hadath}$ corpus: A manually annotated corpus with 550k tokens for event argument relations, used to create $Hadath^{NLI}$ (25k premise-hypotheses pairs).

❖ **Novel Methodology** **for event-argument extraction** by framing it as an NLI problem, achieving high performance.

$Wojood^{OutOfDomain}$: Additional annotated corpus (80k tokens) for **out-of-domain evaluation**.

❖ **End-to-End System** for **event-argument relation extraction**.

Wojood^{OutOfDomain} Corpus

Out of Domain Corpus

❖ **Wojood^{OutOfDomain}** Corpus:

➔ Covers 10 distinct domains.

➔ Years 2010-2022.

➔ 80k tokens.

Out of Domain Results
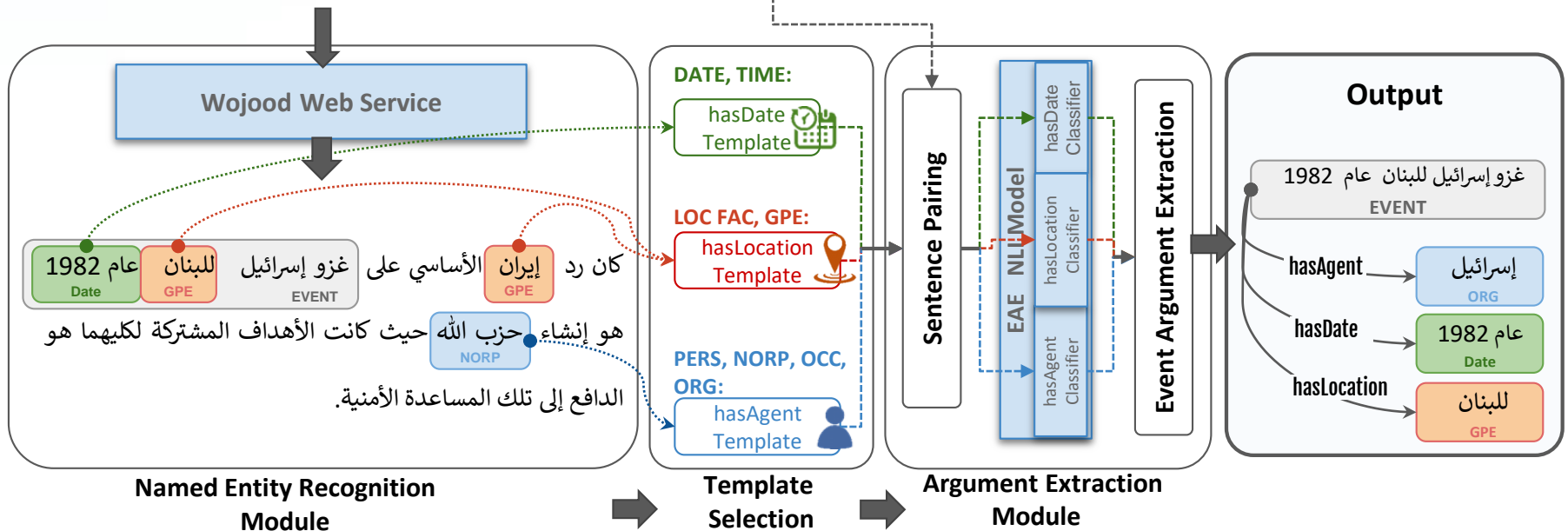
❖ Test the model on **Wojood**^{OutOfDomain}

Class	Support	Precision	Recall	F1 Score
Positive	478	71.05%	78.03%	74.38%
Negative	1,809	94.04%	91.60%	92.80%
Average				**83.59%**

End-to-End Event Relation Extraction

❖ Part of **SinaTools**.

End-to-End
Event Relation Extraction

❖ EAE baselines:

Dataset	Precision	Recall	F1 Score
Hadath^{NLI}	93.45%	94.52%	**93.99%**
Wojood^{OutOfDomain}	67.79%	83.68%	**74.90%**

Thank You

Alaa Aljabari (aaljabari@birzeit.edu)
Lina Duaibes (1205358@student.Birzeit.edu)

References

1. Tymaa Hammouda, Mustafa Jarrar, Mohammed Khalilia: SinaTools: Open Source Toolkit for Arabic Natural Language Understanding. In Proceedings of the 2024 AI in Computational Linguistics (ACLING 2024), Procedia Computer Science, Dubai. ELSEVIER.
2. Mustafa Jarrar, Diyam Akra, Tymaa Hammouda: ALMA: Fast Lemmatizer and POS Tagger for Arabic. In Proceedings of the 2024 AI in Computational Linguistics (ACLING 2024), Procedia Computer Science, Dubai. ELSEVIER.
3. Lina Duaibes, Areej Jaber, Mustafa Jarrar, Ahmad Qadi, Mais Qandeel: Sina at FigNews 2024: Multilingual Datasets Annotated with Bias and Propaganda. In Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024), Bangkok
4. Wajdi Zaghouani, Mustafa Jarrar, Nizar Habash, Houda Bouamor, Imed Zitouni, Mona Diab, Samhaa El-Beltagy, Muhammed AbuOdeh: The FIGNEWS Shared Task on News Media Narratives. In Proceedings of the Second Arabic Natural Language Processing Co
5. Sanad Malaysha, Mo El-Haj, Saad Ezzini, Mohammed Khalilia, Mustafa Jarrar, Sultan Nasser, Ismail Berrada, Houda Bouamor: AraFinNLP 2024: The First Arabic Financial NLP Shared Task. In Proceedings of the Second Arabic Natural Language Processing Con
6. Mustafa Jarrar, Nagham Hamad, Mohammed Khalilia, Bashar Talafha, AbdelRahim Elmadany, Muhammad Abdul-Mageed: WojoodNER 2024: The Second Arabic Named Entity Recognition Shared Task. In Proceedings of the Second Arabic Natural Language Pro
7. Mohammed Khalilia, Sanad Malaysha, Reem Suwaileh, Mustafa Jarrar, Alaa Aljabari, Tamer Elsayed, Imed Zitouni: ArabicNLU 2024: The First Arabic Natural Language Understanding Shared Task. In Proceedings of the Second Arabic Natural Language Processi
8. Alaa Aljabari, Lina Duaibes, Mustafa Jarrar, Mohammed Khalilia: Event-Arguments Extraction Corpus and Modeling using BERT for Arabic. In Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024), Bangkok, Thailand. Asso
9. Mustafa Jarrar, Tymaa Hammouda: Qabas: An Open-Source Arabic Lexicographic Database. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 13363–1337
10. Sanad Malaysha, Mustafa Jarrar, Mohammed Khalilia: NLU-STR at SemEval-2024 Task 1: Generative-based Augmentation and Encoder-based Scoring for Semantic Textual Relatedness In Proceedings of the SemEval 2024 Shared Task 1 (Semantic Relatedness
11. Sylvio Barbon Junior, Paolo Ceravolo, Sven Groppe, Mustafa Jarrar, Samira Maghool, Florence Sèdes, Soror Sahri, and Maurice Van Keulen: Are Large Language Models the New Interface for Data Pipelines? In Proceedings of the International Workshop on Big D
12. Mustafa Jarrar, Sanad Malaysha, Tymaa Hammouda, Mohammed 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.
13. 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 20
14. 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 (Ar
15. Haneen Liqreina, Mustafa Jarrar, Mohammed Khalilia, Ahmed Oumar El-Shangiti, Muhammad AbdulMageed: Arabic Fine-Grained Entity Recognition. In Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023
16. 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 Langua
17. 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 Knowle
18. Nagham Hamad, Mustafa Jarrar, Mohammed 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
19. 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.274-283). San Sebastian, Spain, 2
20. Sanad Malaysha, Mustafa Jarrar, Mohammed Khalilia: Context-Gloss Augmentation for Improving Arabic Target Sense Verification. The 12th International Global Wordnet Conference (GWC2023), Global Wordnet Association. (pp. 254-262). San Sebastian, Spain,
21. 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
22. 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 Ap
23. 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
24. Mustafa Jarrar: The Arabic Ontology - An Arabic Wordnet with Ontologically Clean Content. Applied Ontology Journal, 16:1, 1-26. IOS Press. 2021
25. 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
26. 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 Cr
27. 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 Comp
28. 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
29. 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 Languag
30. Mustafa Jarrar: Digitization of Arabic Lexicons. Arabic Language Status Report. UAE Ministry of Culture and Youth. Pages 214-2017. Dec 2020
31. 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
32. 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. 201
33. 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
34. 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
35. Paolo Ceravolo, Antonia Azzini, Marco Angelini, Tiziana Catarci, Philippe Cudre-Mauroux, Ernesto Damiani, Alexandra Mazak, Maurice Van Keulen, Mustafa Jarrar, Giuseppe Santucci, Kai-Uwe Sattler, Monica Scannapieco, Manuel Wimmer, Robert Wrembel, Fad
36. 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
37. Mustafa Jarrar, Werner Ceusters: Classifying Processes and Basic Formal Ontology. The 8th International Conference on Biomedical Ontology (ICBO), Newcastle, UK. September, 2017
38. 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
39. 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), Issue(3). Springer (doi.org/10.1007/s10579-01
40. Mamoun Abu Helou, Matteo Palmonari, Mustafa Jarrar: Effectiveness of Automatic Translations for Cross-Lingual Ontology Mapping. Journal of Artificial Intelligence Research, Special Track on Cross-language Algorithms and Applications. Pages(165-208). Volume
41. Mustafa Jarrar, Anton Deik: The Graph Signature: A Scalable Query Optimization Index for RDF Graph Databases Using Bisimulation and Trace Equivalence Summarization. International Journal on Semantic Web and Information Systems, 11(2), Pages(36-65). IG
42. Mustafa Jarrar, Nizar Habash, Diyam Akra, Nasser Zalmout: Building a Corpus for Palestinian Arabic: a Preliminary Study. Arabic Natural Language Processing Workshop, at the Conference on Empirical Methods in Natural Language Processing (EMNLP 2014).
43. Mamoun Abu Helou, Matteo Palmonari, Mustafa Jarrar, Christiane Fellbaum: Towards Building Lexical Ontology via Cross-Language Matching. The 7th Conference on Global WordNet. Pages(346-354). Global WordNet Association. ISBN:7492329944978. Tartu, E
44. Antonio Lucas Soares, Carla Sofia Pereira, Mustafa Jarrar (eds): Proceedings of the International Workshop on Ontology Content and Evaluation (OnToContent 2014). In OTM 2014 Workshops. Page 575. LNCS 8842, Springer. ISBN:9783662455494. October, 201
45. Mustafa Jarrar and Marios D. Dikaiakos: A Query Formulation Language for the Data Web. The IEEE Transactions on Knowledge and Data Engineering. Volume 24, Number 4, Pages(783-798). IEEE Computer Society. April, 2012
46. Gianluca Elia, Mustafa Jarrar: Guest Editorial: Knowledge Management and e-Human Resources Practices for Innovation. The International Journal of Knowledge and Learning (IJKL). Pages(1-5). Volume (8), Number(1/2). Inderscience Publishers. 2012
47. Mustafa Jarrar, Amanda Hicks, Matteo Palmonar (eds): Proceedings of the International Workshop on Ontology Content and Evaluation (OnToContent 2012) . In OTM 2012 Workshops. Page 419. LNCS 7567, Springer. ISBN:9783642336171. September, 2012
48. Mustafa Jarrar: Proceedings of the 1st Palestinian Conference on e-Governance and e-Services. Sina Institute at Birzeit University. June, 2012
49. Mustafa Jarrar: Building a Formal Arabic Ontology (Invited Paper). Proceedings of the Experts Meeting on Arabic Ontologies and Semantic Networks. ALECSO, Arab League. Tunisia. July, 2011
50. Mustafa Jarrar, Anton Deik, Bilal Faraj: Ontology-Based Data and Process Governance Framework -The Case of e-Government Interoperability in Palestine. The IFIP International Symposium on Data-Driven Process Discovery and Analysis (SIMPDA). Pages(83-9
51. Paolo Ceravolo, Chengfei Liu, Mustafa Jarrar, Kai-Uwe Sattler: Special Issue on Querying the Data Web -Novel techniques for querying structured data on the web. The World Wide Web Journal. Volume(14), Issue (5-6). Springer. ISSN:1573-1413. August, 201