ArabicNLU 2024:

The 1st Arabic Natural Language Understanding Shared Task


Mohammed Khalilii


Sanad Malaysha


Reem Suwaileh


Mustafa Jarrar


Alaa Aljabari


Tamer Elsayed


Imed Zitouni

BIRZEIT UNIVERSITY

Google

جامعة قطر
QATAR UNIVERSITY

جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

Open Source

Arabic NLP Tools and Datasets

جامعة بيرزيت
BIRZEIT UNIVERSITY

https://sina.birzeit.edu/resources

SinaLab

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

- **Natural Language Understanding is a fundamental** aspect of NLP.
- Enable **semantic-based** human-computer interactions.
- A major **challenge in Arabic** due to its morphological richness.

ArbicNLU2024 Shared Task

Subtask 1 Word Sense Disambiguation

Best: 78%

Baseline: 84%

SALMA Corpus

34k sense-annotated tokens

Subtask 2 Location Mention Disambiguation

Best: 95%

Baseline: 62.7%

IDRISI-DA Corpus

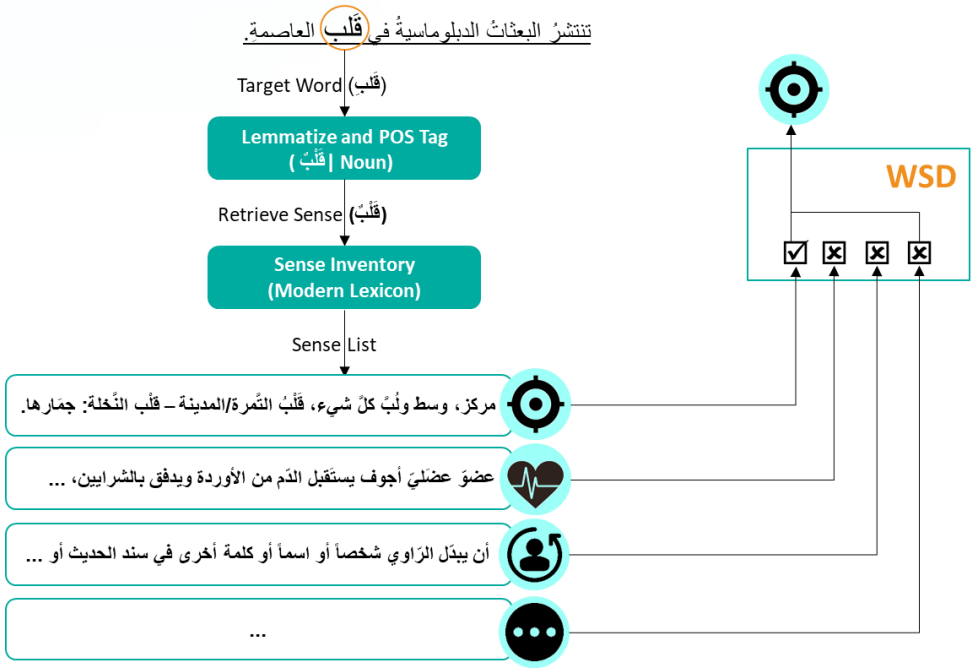
3900 annotations

Natural Language Understanding

- **Subtask-1 Word Sense Disambiguation (WSD):** to disambiguate the semantics of polysemous words.
- **Subtask-2 Location Mention Disambiguation (LMD):** to resolve location mentions and linking them to toponyms in geo-databases.

Task Description

Subtask1 : Word Sense Disambiguation (WSD):



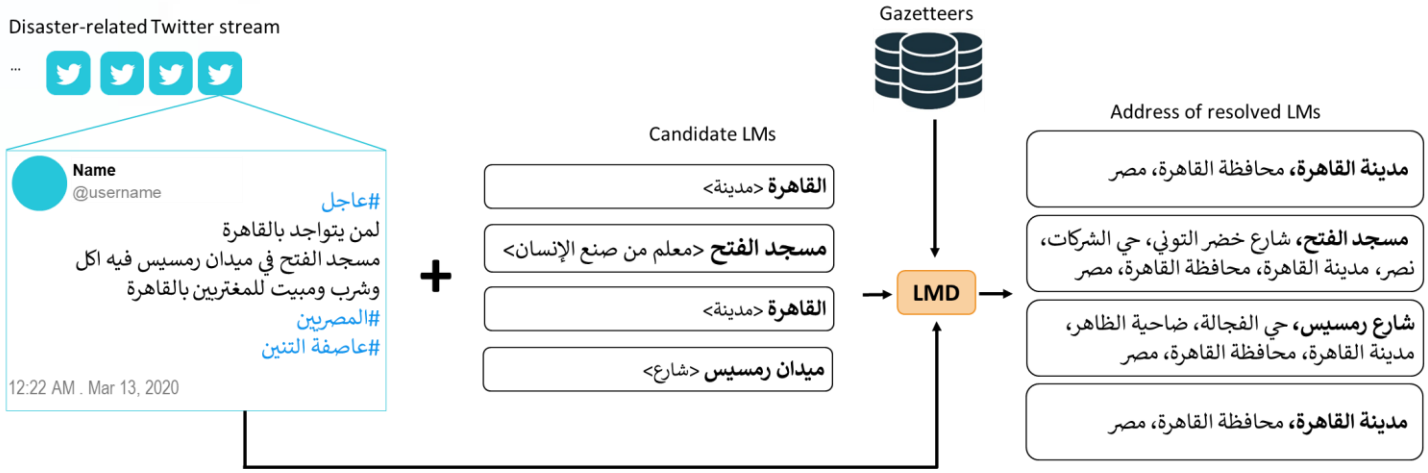
WSD Dataset (SALMA)

First sense-annotated corpus for Arabic:

- 1,440 sentences.
- 34K tokens 8,760 unique tokens with 3,875 unique lemmas).
- A total of 4,151 senses.

Task Description

Subtask2 : Location Mention Disambiguation (LMD):

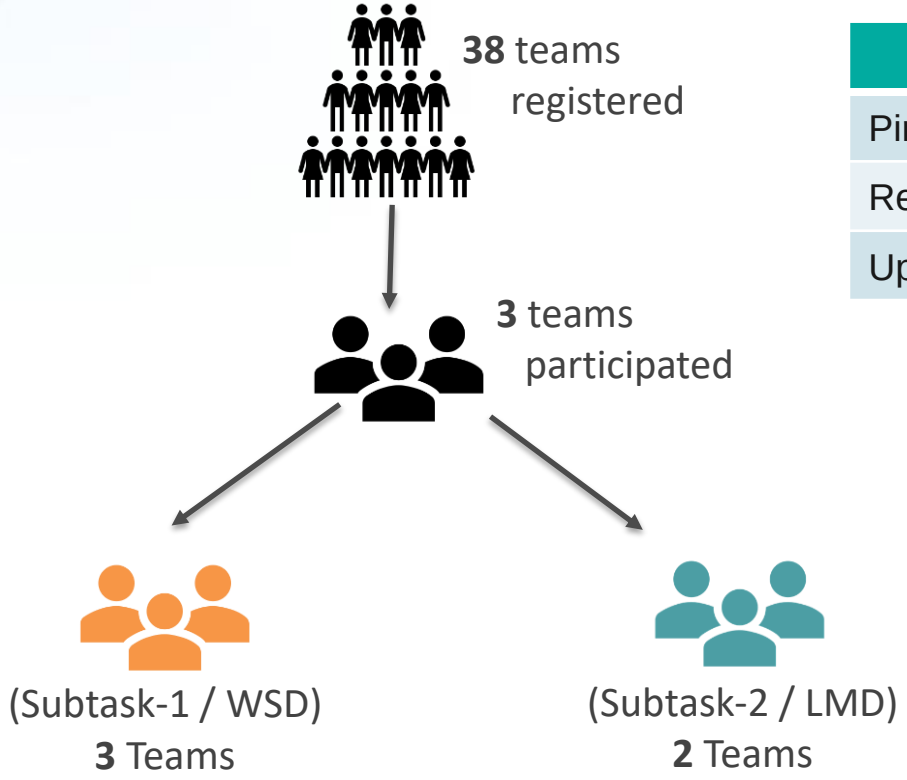


LMD Dataset (IDRISI)

First Arabic LMD dataset. It contains:

- 2,869 posts from diverse dialects.
- 3,893 location mentions, of which 763 are unique, across seven countries.

Shared Task Teams & Results



Team	Affiliation	Task
Pirates	Nile University	1
Rematchka	Cairo University	1, 2
Upaya	SCB DataX	1, 2

Participating Systems

Subtask-1 WSD (baseline 84.2%):

- **Top team** (77.8%) used Llama3-70B-Instruct for zero-shot learning, outperforming GPT-4.
- **2nd team** (77.8%) fine-tuned AraBERTv2 as a sense classifier.
- **3rd team** (57.5%) employed zero-shot learning using LLMs and finetuning PLMs.

Subtask-2 Location Mention Disambiguation (LMD) (baseline 57.24%): **:**

- **Top team** (94.97%) used Llama3 for translation and GeoPy for toponym retrieval.
- **2nd team (**59.94%) used Open Street Map for retrieval and Cohere for reranking.

Shared Task Teams & Results

Results of Subtask 1 – WSD

Rank	Team	Accuracy
	Baseline	**84.2%**
1	Upaya	77.8%
2	Pirates	70.8%
3	Rematchka	57.5%

Results of Subtask 2 – LMD

Rank	Team	MRR@1	MRR@2	MRR@3
1	Rematchka	94.97%	95.00%	95.00%
2	Upaya	59.94%	59.94%	59.94%
	Baseline	**57.24%**	**63.96%**	**64.28%**

Open Challenges

- **WSD and LMD are challenging tasks!!** Need for more datasets and more research.
- **Arabic dialectics** should be supported in WSD and LMD.
- **LLMs performed badly** - compared to classification architectures.
- Maybe **Arabic-tailored LLMs are needed**!!.

Thank You

Alaa Aljabari

aaljabari@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 2
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 Languag
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 Knowled
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). P
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:7492329949978. 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