The First Arabic Natural Language Processing Conference (ArabicNLP 2023)
Singapore
7/12/2023

ArBanking77: Intent Detection Neural Model and a New Dataset in Modern and Dialectical Arabic

Mustafa Jarrar
Birzeit University, Palestine

Ahmet Birim
Sestek, Türkiye

Mohammed Khalilia
Birzeit University, Palestine

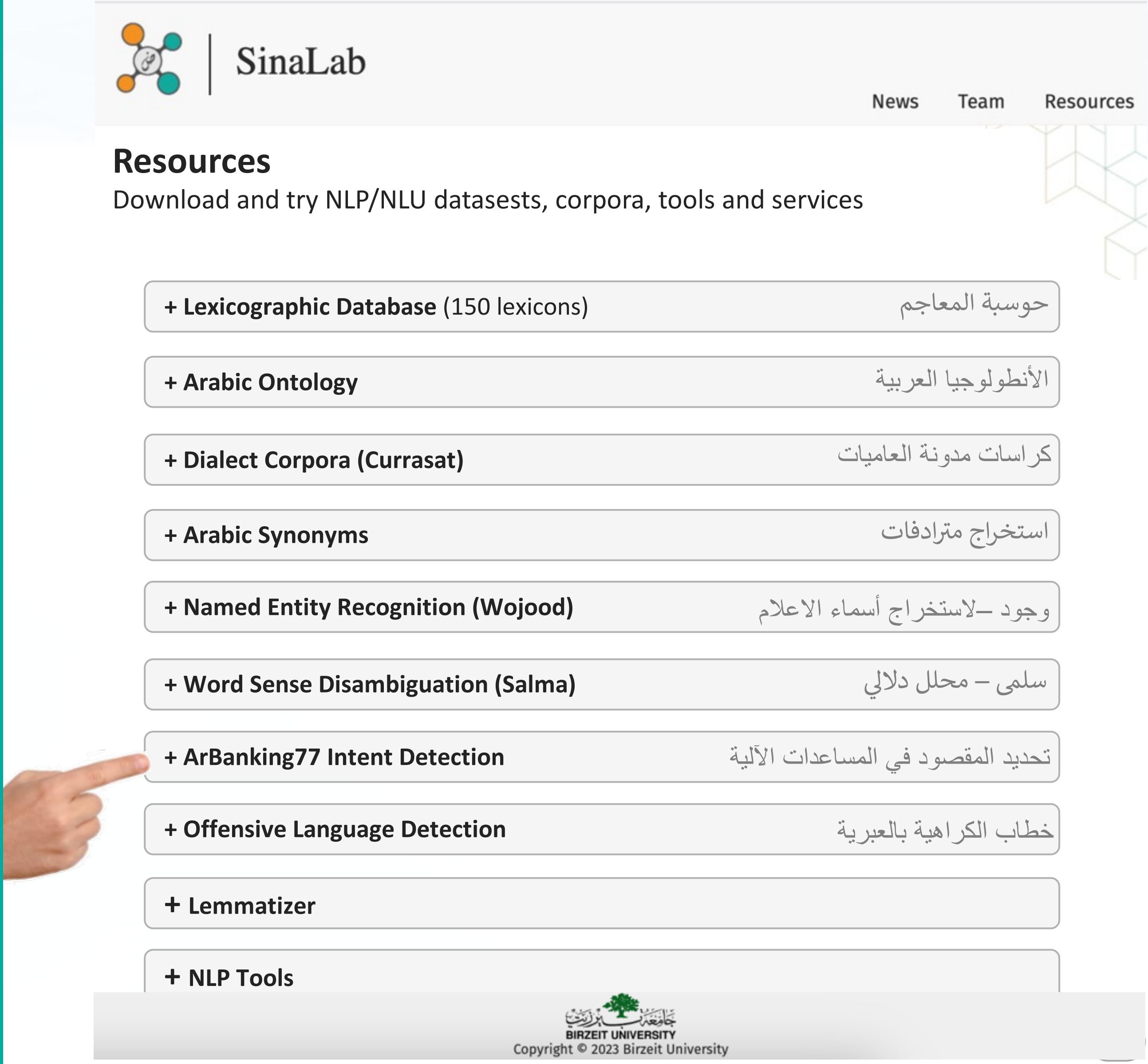
Mustafa Erden
Sestek, Türkiye

Sana Ghanem
Birzeit University, Palestine

Natural Language Understanding Tools and Datasets

Open Source

https://sina.birzeit.edu/resources



SinaLab

News Team Resources

Resources

Download and try NLP/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

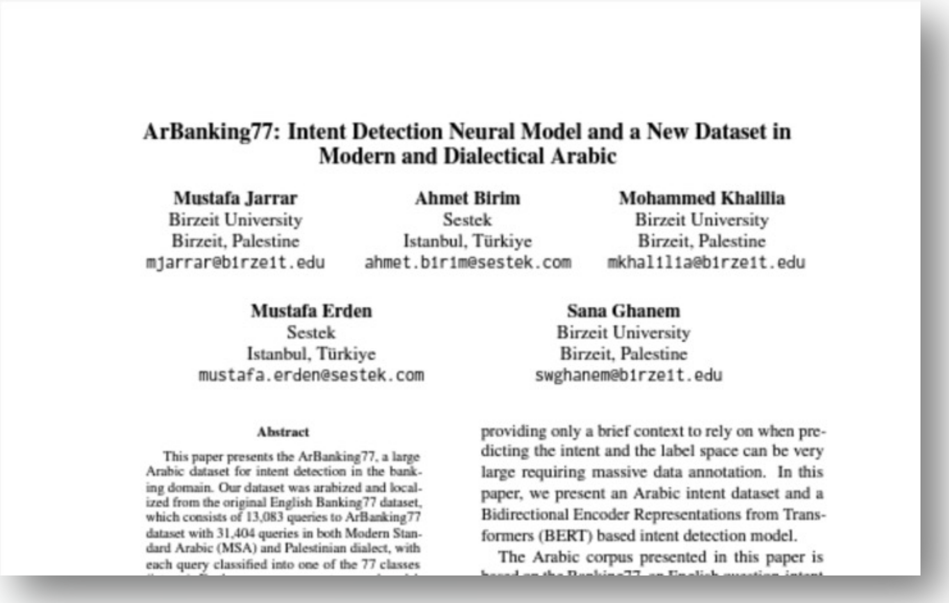
+ **ArBanking77 Intent Detection** تحديد المقصود في المساعدات الآلية

+ **Offensive Language Detection** خطاب الكراهية بالعبرية

+ **Lemmatizer**

+ **NLP Tools**

جامعة بيرزيت
BIRZEIT UNIVERSITY
Copyright © 2023 Birzeit University



ArBanking77: Intent Detection Neural Model and a New Dataset in Modern and Dialectical Arabic

Mustafa Jarrar
Birzeit University
Birzeit, Palestine
mjarrar@birzeit.edu

Ahmet Birim
Sestek
Istanbul, Türkiye
ahmet.birim@sestek.com

Mohammed Khalilia
Birzeit University
Birzeit, Palestine
mkhalilia@birzeit.edu

Mustafa Erden
Sestek
Istanbul, Türkiye
mustafa.erden@sestek.com

Sana Ghanem
Birzeit University
Birzeit, Palestine
swghanem@birzeit.edu

Abstract

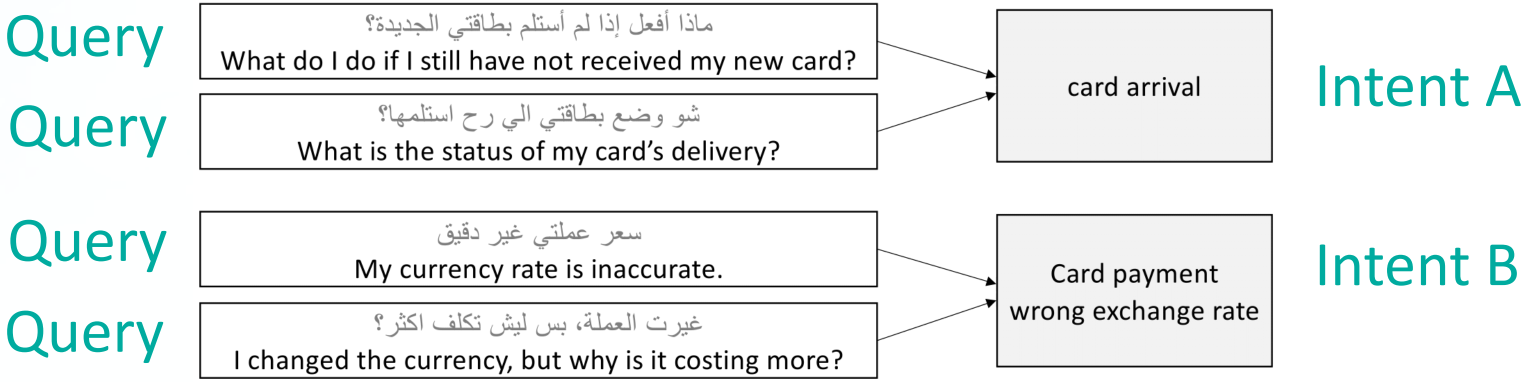
This paper presents the ArBanking77, a large Arabic dataset for intent detection in the banking domain. Our dataset was arabized and localized from the original English Banking77 dataset, which consists of 13,083 queries to ArBanking77 dataset with 31,404 queries in both Modern Standard Arabic (MSA) and Palestinian dialect, with each query classified into one of the 77 classes

providing only a brief context to rely on when predicting the intent and the label space can be very large requiring massive data annotation. In this paper, we present an Arabic intent dataset and a Bidirectional Encoder Representations from Transformers (BERT) based intent detection model.

The Arabic corpus presented in this paper is

Jarrar, M., Birim, A., Khalilia, M., Erden, M., Ghanem, S. (2023) ArBanking77: Intent Detection Neural Model and a New Dataset in Modern and Dialectical Arabic. In Proceedings of the Arabic Natural Language Processing Conference (ArabicNLP 2023), Singapore.

The problem



Query

ماذا أفعل إذا لم أستلم بطاقتي الجديدة؟

What do I do if I still have not received my new card?

Query

شو وضع بطاقتي الي رح استلمها؟

What is the status of my card’s delivery?

card arrival

Intent A

Query

سعر عملتي غير دقيق

My currency rate is inaccurate.

Query

غيرت العملة، بس ليش تكلف اكثر؟

I changed the currency, but why is it costing more?

Card payment wrong exchange rate

Intent B

Challenging to: build an Arabic intent dataset, and train BERT.

Our Contributions

❖ArBanking77 dataset

- ArBanking77 dataset consists of 31,404 queries.
- 2.4x larger than the Banking77 dataset.
- On average, there are 408 queries per intent
  - 202 MSA queries/intent
  - 206 Palestinian queries/intent.

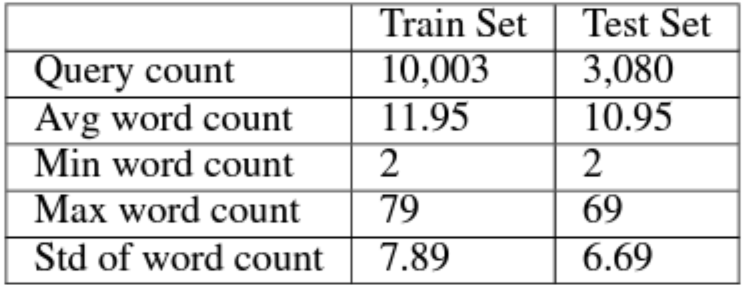
❖Intent detection model

- AraBERTv2
- F1-scores on MSA and PAL are 0.9209 and 0.8995, respectively

Corpus Collection

The ArBanking77 corpus:

- Derived from the Banking77 dataset
  - 13,083 queries
  - 77 classes (intents)
  - Single domain, banking
  - Open under the (CC-BY-4.0) license
  - The original Banking77 dataset is divided into train and test dataset.



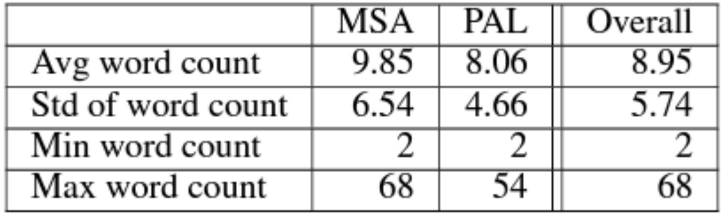
	Train Set	Test Set
Query count	10,003	3,080
Avg word count	11.95	10.95
Min word count	2	2
Max word count	79	69
Std of word count	7.89	6.69

Corpus Collection

The ArBanking77 corpus:

Each query in the original Banking77 has at least two corresponding queries in the ArBanking77

- At least one query written in MSA.
- At least one query written in Palestinian dialect.



	MSA	PAL	Overall
Avg word count	9.85	8.06	8.95
Std of word count	6.54	4.66	5.74
Min word count	2	2	2
Max word count	68	54	68

Annotation Process

- 26 annotators (Well trained)
- Done using Google Sheets
- Over several months

- Phases:

  Phase I: Arabization and Localization

  Phase II: Review

Annotation Process

Phase I: Arabization and Localization

1. The translation of the Banking77 from English into MSA.
   - Done using Google Translate API.

2. The manual annotation .

Annotation Process

Phase I: Arabization and Localization

The annotators performed four steps for each original English query:

(i) MSA_1 should be revised in case of incorrect translation.
(ii) MSA_2 is optionally written by the annotator.
(iii) PAL_1 is the formulation of the query in the Palestinian dialect.
(iv) PAL_2 is optionally written by the annotator.

Each intent was divided among 2-5 annotators.

Annotation Process

Phase II: Review

Step1: Each annotator reviewed three related intents, to ensure that:

(i) The MSA and Palestinian queries should be acceptable, semantically correct and well-formulated.

(ii) All queries in one intent belong to that intent, and not to other intents (labeling consistency).

(iii) Spelling mistakes are ignored in order to simulate common errors and noise in real NLP systems, especially in live chat queries.

Step2: We revised duplicate queries by introducing additional variations to make them unique.

Final Dataset

Lexical Relation between MSA and PAL

- Measured using the Jaccard Index for each parallel pair (MSA and PAL)

  Results of Jaccard index:

  - The mean is 0.16,
  - The median 0.13
  - The standard deviation 0.13.

- Thus, for diaglossic languages such as Arabic, training on one variation is not necessarily extensible.

Intent Detection Model

Transformer-based Intent Classifier

- BERT encoder is fine-tuned on Arabic intent detection task using the ArBanking77 dataset.
- A single linear layer was added on top of BERT transformer layers to perform the intent classification task.

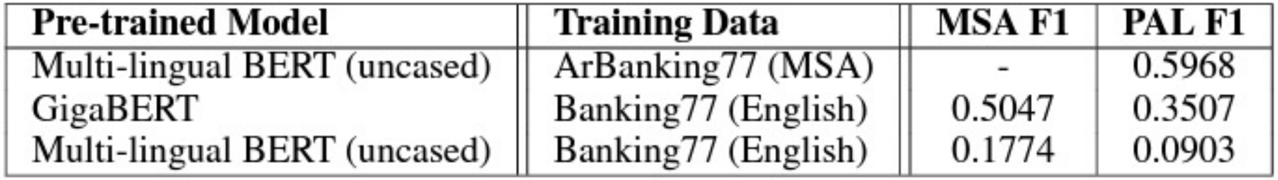
Model Training

- Training (21,559 queries), validation (2,464 queries) and test (7,381 queries).
- Learning rate, $\eta = 4e^{-5}$
- Batch size of 64, maximum of 20 epochs

Experiments and Results

Zero-Shot Cross-Lingual Transfer Learning

- Used multi-lingual BERT (mBERT) (Devlinetal.,2018) and GigaBERT (Lanetal., 2020).



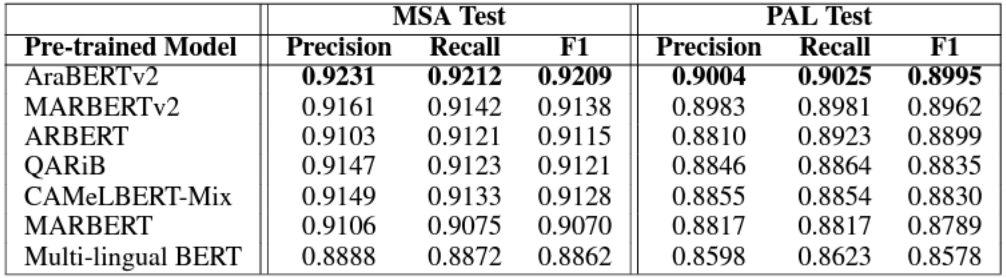
Pre-trained Model	Training Data	MSA F1	PAL F1
Multi-lingual BERT (uncased)	ArBanking77 (MSA)	-	0.5968
GigaBERT	Banking77 (English)	0.5047	0.3507
Multi-lingual BERT (uncased)	Banking77 (English)	0.1774	0.0903

- Multilingual pre-trained transformers did not perform well on MSA and PAL.

Experiments and Results

Pre-Trained Transformers Benchmark

- ❑ Evaluate various Arabic pre-trained transformer models, we benchmark against these models:



	MSA Test			PAL Test		
Pre-trained Model	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
AraBERTv2	**0.9231**	**0.9212**	**0.9209**	**0.9004**	**0.9025**	**0.8995**
MARBERTv2	0.9161	0.9142	0.9138	0.8983	0.8981	0.8962
ARBERT	0.9103	0.9121	0.9115	0.8810	0.8923	0.8899
QARiB	0.9147	0.9123	0.9121	0.8846	0.8864	0.8835
CAMeLBERT-Mix	0.9149	0.9133	0.9128	0.8855	0.8854	0.8830
MARBERT	0.9106	0.9075	0.9070	0.8817	0.8817	0.8789
Multi-lingual BERT	0.8888	0.8872	0.8862	0.8598	0.8623	0.8578

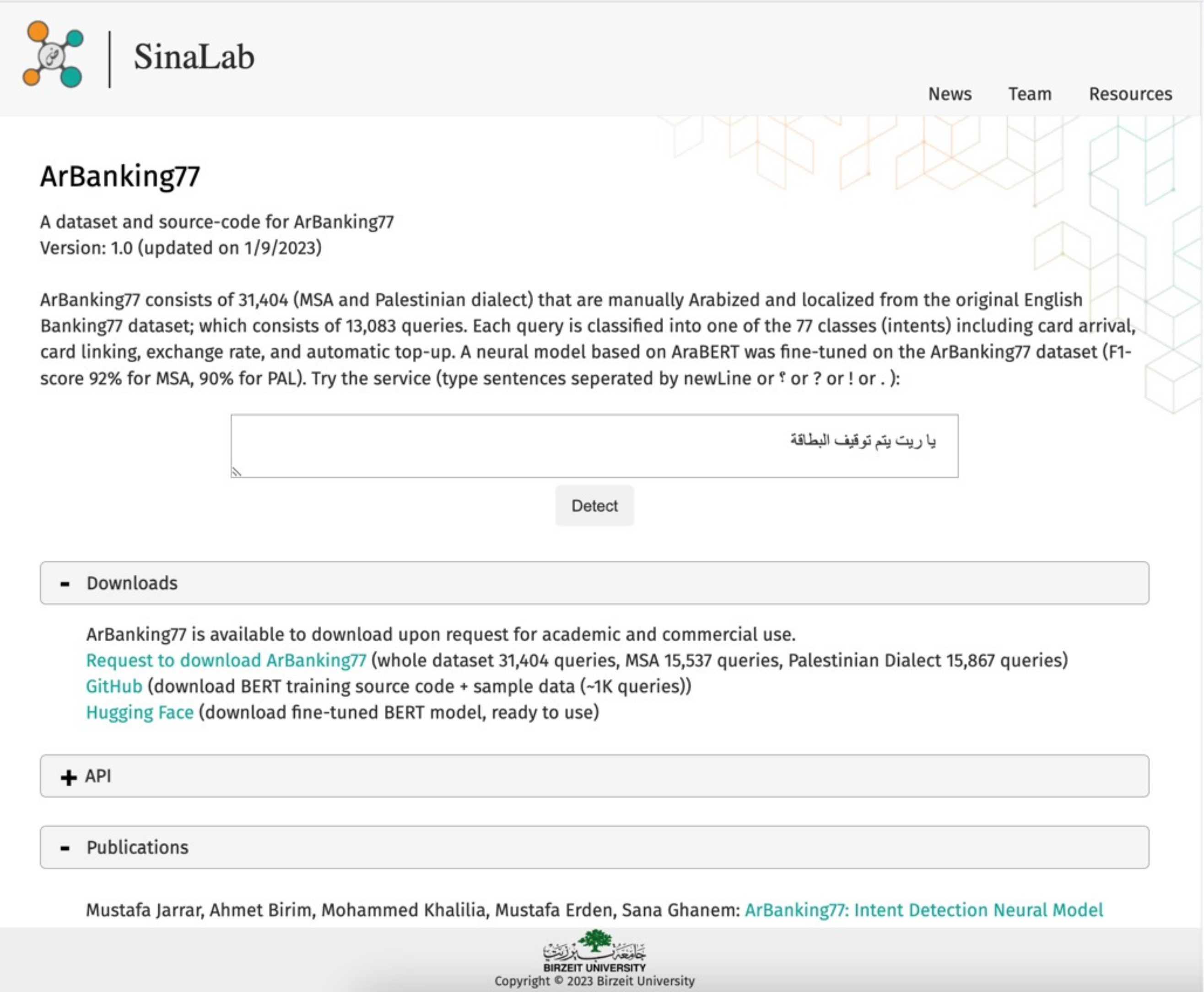
Results:

AraBERTv2 gives the best F1-score.

Summary

- ArBanking77 is the first Arabic intent detection dataset in the banking domain.
- Benchmarked different models for intent detection.
- AraBERTv2 is the best model for Arabic dialectical dataset.

Download



ArBanking77

A dataset and source-code for ArBanking77
Version: 1.0 (updated on 1/9/2023)

ArBanking77 consists of 31,404 (MSA and Palestinian dialect) that are manually Arabized and localized from the original English Banking77 dataset; which consists of 13,083 queries. Each query is classified into one of the 77 classes (intents) including card arrival, card linking, exchange rate, and automatic top-up. A neural model based on AraBERT was fine-tuned on the ArBanking77 dataset (F1-score 92% for MSA, 90% for PAL). Try the service (type sentences seperated by newLine or ؟ or ? or ! or .):

يا ريت يتم توقيف البطاقة

- Downloads

ArBanking77 is available to download upon request for academic and commercial use.
Request to download ArBanking77 (whole dataset 31,404 queries, MSA 15,537 queries, Palestinian Dialect 15,867 queries)
GitHub (download BERT training source code + sample data (~1K queries))
Hugging Face (download fine-tuned BERT model, ready to use)

+ API

- Publications

Mustafa Jarrar, Ahmet Birim, Mohammed Khalilia, Mustafa Erden, Sana Ghanem: ArBanking77: Intent Detection Neural Model

https://sina.birzeit.edu/arbanking77/

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).