



Arabic Dialects Papers

Dr. Hajer

Student Name: Amro Ismail Mohamed (Papers 1 to 3)

Student Name: Mohamed Alghazwani (Papers 4 to 6)

1. Summary of the Research Papers

Paper 1:

NADI 2024: The Fifth Nuanced Arabic Dialect Identification Shared Task

The paper for NADI 2024 and its objective is to help advance SoTA Arabic NLP by providing guidance, datasets, modeling opportunities, and standardized evaluation conditions that allow researchers to compete on prespecified tasks collaboratively. They created three subtasks, targeting both dialect identification, cast as a multi-label task (Subtask 1), identification of the Arabic level of dialectness (Subtask 2), and dialect-to-MSA machine translation (Subtask 3). The results indicate that there is still room for improvement across the various tasks. In the future, they intend to cover more Arabic dialects in NADI and propose novel ways of modeling that utilize large language models.

Paper 2:

Alexandria: A Multi-Domain Dialectal Arabic Machine Translation Dataset for Culturally Inclusive and Linguistically Diverse LLMs

The paper for Alexandria focusing on serves as both a training resource and a rigorous benchmark for evaluating MT and Large Language Models (LLMs). It covers 13 Arab countries and 11 high-impact domains, including health, education, and agriculture. The authors' evaluations reveal critical gaps in existing models regarding regional dialects, technical terminology, and gender alignment. By releasing Alexandria, they provide a robust framework to address these limitations, fostering the development of more accurate and culturally sensitive language technologies.

Paper 3:

AraT5: Text-to-Text Transformers for Arabic Language Generation

The paper focuses on the investigation of how the Transformer framework (T5), which converts all language problems into a text-to-text format, was recently proposed as a simple and effective transfer learning approach. Although a multilingual version of the T5 model (mT5) was also introduced, it is not clear how well it can fare on non-English tasks involving diverse data. So the authors applied mT5 on a language with a wide variety of dialects—Arabic. For evaluation, they introduced a novel benchmark for **AR**abic language **GEN**eration (ARGEN), covering seven important tasks. For model comparison, they pre-trained three powerful Arabic T5-style models and evaluated them on ARGEN. Although pre-trained with $\sim 49\%$ less data, their new models perform significantly better than mT5 on all ARGEN tasks (in 52 out of 59 test sets) and set several new SOTAs.

Paper	All Models	Datasets	Limitations	Best Model	Results
NADI 2024: The Fifth Nuanced Arabic Dialect Identification Shared Task	MARBERT / MARBERTv2 AraBERT mMARCO Sentence ALDi Model AraT5 mT5 mBART / mBART-50	LAHJATBERT Repository [https://mohamedalaa9.github.io/lahjatbert/] MADAR Parallel Corpus [https://sites.google.com/nyu.edu/madar/home#h.xpcfdhjyc95c]	This edition of NADI focused on only 10 country-level dialects for Subtasks 1 and 2 NADI continues to use short texts for the Arabic dialects limited resources	Subtask1: fine-tuned BERT-based models Subtask2: MARBERT model modified with a continuous regression head Subtask3: AraT5 and mT5 architectures	Subtask 1 (F: 50.57) Subtask 2 (RMSE: 0.1403) Subtask 3 (BLEU: 20.44)
Alexandria: A Multi-Domain Dialectal Arabic Machine Translation Dataset for Culturally Inclusive and Linguistically Diverse LLMs	NLLB-200-3.3B gemma-3-12b-it aya-expanse-32b c4ai-command-a-03-2025	Alexandria Dataset [https://huggingface.co/datasets/UBC-NLP/alexandria]	Gender Imbalance in Scenarios Technical Lexical Gaps and MSA Leakage Restricted Closed-Models Evaluation	c4ai-command-a-03-2025	Eg: 29.45 Sa: 29.4
AraT5: Text-to-Text Transformers for Arabic Language Generation	mT5 AraT5Tw AraT5MSA AraT5	AraT5 [https://github.com/UBC-NLP/araT5] [https://huggingface.co/UBC-NLP/AraT5v2-base-1024]	data quality/cleanliness challenges limit resources	AraT5MSA	AraT5MSA is at 65.88 BLEU

References

- ✚ [NADI 2024: The Fifth Nuanced Arabic Dialect Identification Shared Task](<https://aclanthology.org/2024.arabnlp-1.79/>) (Abdul-Mageed et al., ArabicNLP 2024)
- ✚ Alexandria: A Multi-Domain Dialectal Arabic Machine Translation Dataset for Culturally Inclusive and Linguistically Diverse LLMs}, <https://arxiv.org/abs/2601.13099>
- ✚ [AraT5: Text-to-Text Transformers for Arabic Language Generation](<https://aclanthology.org/2022.acl-long.47/>) (Nagoudi et al., ACL 2022)

Paper 4:**MARASTA: A Multi-dialectal Arabic Cross-domain Stance Corpus**

Charfi et al. introduced MARASTA, a cross-domain and multi-dialectal Arabic stance corpus. The corpus contains 4,657 annotated sentences collected from several Arabic regions and covers different controversial topics. The dataset was used to evaluate various machine learning models for stance detection. Stance detection is an important NLP task used to determine whether a text supports, opposes, or is neutral toward a topic. The increasing use of social media has made stance detection valuable for opinion analysis and polarization studies. Previous Arabic stance datasets focused on limited topics or specific dialects. Existing resources often suffered from class imbalance and limited dialectal coverage. The authors collected data from Twitter and YouTube. Sentences were manually annotated into Favor, Against, and Neutral classes. Several machine learning experiments were conducted using unigram and TF-IDF features. The Multi-Layer Perceptron (MLP) model achieved the best performance with an Accuracy of 0.66 and an F1-score of 0.66. MARASTA provides a balanced and multi-dialectal Arabic corpus that can support future research in Arabic stance detection.

Paper 5:**NADI 2022: The Third Nuanced Arabic Dialect Identification Shared Task**

The paper presents NADI 2022, a shared task that focuses on Arabic dialect identification and dialectal sentiment analysis. It provides benchmark datasets and evaluation procedures. Arabic dialect identification remains challenging because of the large variation among dialects. Benchmark tasks help researchers compare different approaches. The paper reviews previous NADI shared tasks and several Arabic dialect resources and sentiment analysis datasets. Two subtasks were proposed: dialect identification and dialect sentiment analysis. Participants used Twitter datasets and machine learning methods to solve both tasks. The best systems achieved an F1-score of 27.06 for dialect identification and 75.16 for sentiment analysis. NADI 2022 provides a benchmark that encourages future research in Arabic NLP and dialect processing.

Paper 6:**ADI-20: Arabic Dialect Identification Dataset and Models**

ADI-20 extends ADI-17 and introduces a larger Arabic speech dataset covering 20 dialects and Modern Standard Arabic. Dialect identification is essential for speech recognition and language processing applications. Arabic dialects show large linguistic variations. Previous ADI datasets and shared tasks provided speech resources for Arabic dialect identification, but some dialects were underrepresented. The authors investigated ECAPA-TDNN and Whisper encoder architectures. They also studied the effect of training data size on performance. The experiments showed that using only 30% of the training data caused only a slight reduction in performance. ECAPA-TDNN achieved strong results. ADI-20 improves dialect coverage and provides publicly available resources for future research.

Paper	All Models	Datasets	Limitations	Best Model	Results
MARASTA: A Multi-dialectal Arabic Cross-domain Stance Corpus	MLP	MARASTA	Previous datasets were unbalanced	MLP	Accuracy=0.66, F1=0.66
NADI 2022: The Third Nuanced Arabic Dialect Identification Shared Task	Transformer-based models	NADI datasets	Dialect complexity	Winning System	F1=75.16
ADI-20: Arabic Dialect Identification Dataset and Models	ECAPA-TDNN, Whisper	ADI-20	Limited dialect resources	ECAPA-TDNN	High F1-score

References

- [MARASTA: A Multi-dialectal Arabic Cross-domain Stance Corpus](<https://aclanthology.org/2024.lrec-main.964/>) (Charfi et al., LREC-COLING 2024)
- [NADI 2022: The Third Nuanced Arabic Dialect Identification Shared Task](<https://aclanthology.org/2022.wanlp-1.9/>) (Abdul-Mageed et al., WANLP 2022)
- Elleuch, H., et al. (2025). ADI-20: Arabic Dialect Identification Dataset and Models[<https://arxiv.org/abs/2511.10070>]