Arabizi-Arabic Neural Transliteration System

Problem

The Arabizi Problem

- Arabic script is not compatible with standard keyboards
- Social media users use Roman letters (Arabizi)
- Romanization depends on local dialect and colonial language

Intelligence & Security Implications

- OSINTers struggle to search Arabic social media
- Terrorist/hate speech monitoring misses Arabizi content
- LE cannot query Arabic databases with Romanized input

Current Tools Gap

- Existing transliteration tools are rule-based and inaccurate
- No dialect aware bidirectional neural approaches conversion tools
- Need for real-time, accurate transliteration at scale

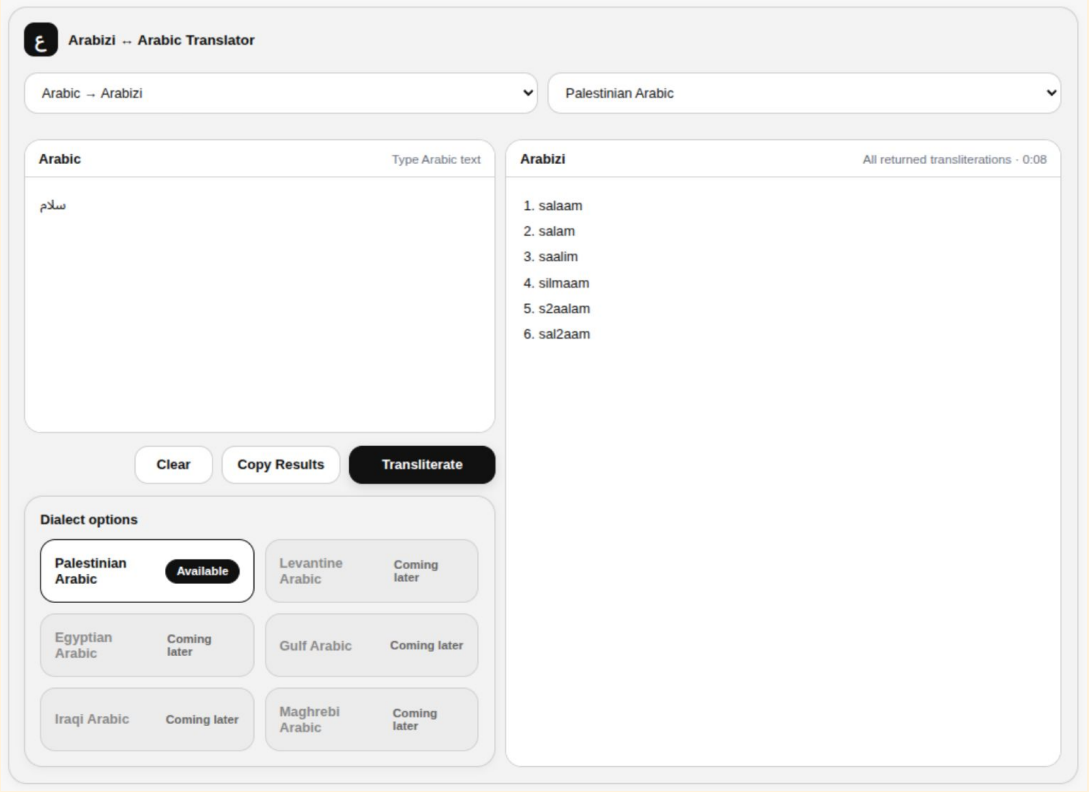
Solution

Arabizi-Arabic Transliterator

- A Transformer-based neural transliteration system
- Bidirectional conversion: Arabizi ↔ Arabic
- Real-time web/API demo with top-k outputs

Key Features

- Separate sequence-to-sequence models per direction
- Beam search decoding for stronger candidate recall
- Learned MLP re-ranker after beam generation
- Unicode normalization, diacritic handling, and Arabizi digit tokens
- Returns top-12 predictions with beam/reranker scores



High-Level Architecture

Architecture Components

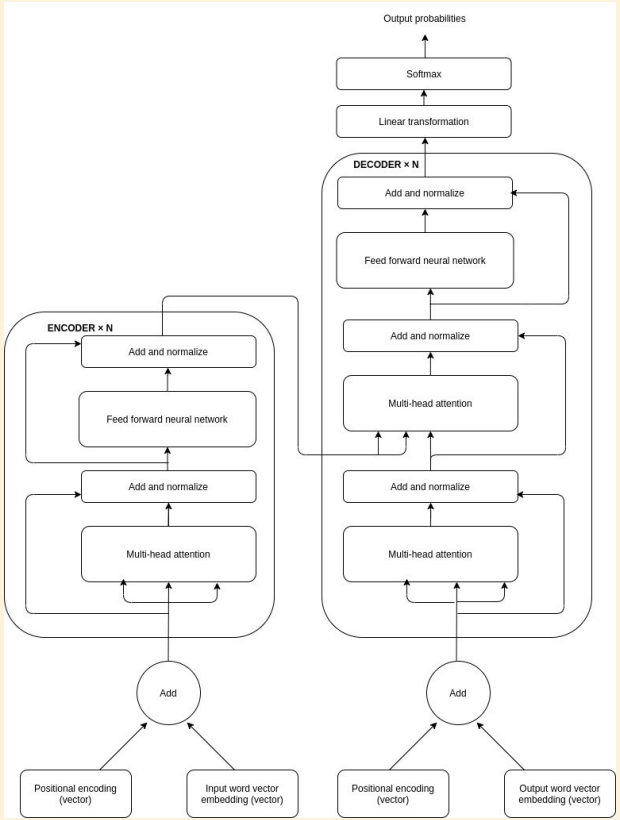
- Transformer encoder-decoder models
- Encoder: 4 layers, 8 attention heads
- Decoder: 4 layers, 8 attention heads
- Embedding dimension: 256
- Feed-forward dimension: 1024
- Dropout: 15%

Input Processing

- Direction-specific tokenization
- Arabic: character-level tokens
- Arabizi: character + multi-token patterns like sh, kh, 3a, 7o
- Special tokens: <PAD>, <BOS>, <EOS>, <UNK>
- Unicode normalization (NFC) + lowercase Arabizi

Output Generation

- Beam search width=20 with length penalty
- Returns top-12 predictions per input



Architecture Deep Dive

Vocabulary

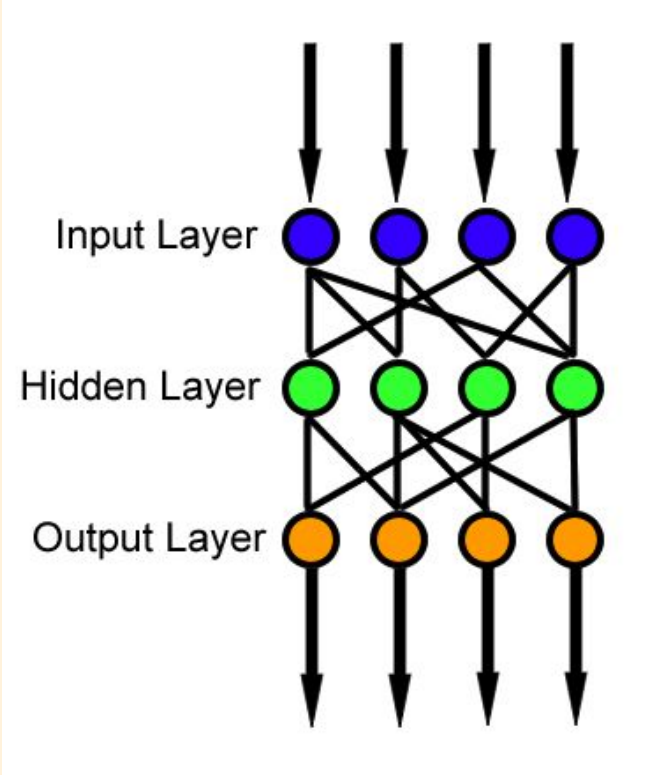
- Direction-specific source/target vocabularies
- Arabic vocab: 44 tokens
- Arabizi vocab: 54–65 tokens depending on direction
- Covers Arabic letters/diacritics, Roman letters, digits, spaces, and common digraphs

Training Configuration

- Loss: cross-entropy with 0.02 label smoothing
- Optimizer: AdamW with cosine learning-rate schedule
- Batch size: 96; max 100 epochs; early stopping patience 16
- Length curriculum + capped weighted sampling
- Checkpoints and metrics autosaved to Google Drive

Deployment Artifacts

- Base model website bundles include tokenizer, weights, vocab, and config
- Re-ranker .pt files are saved separately and loaded with the backend



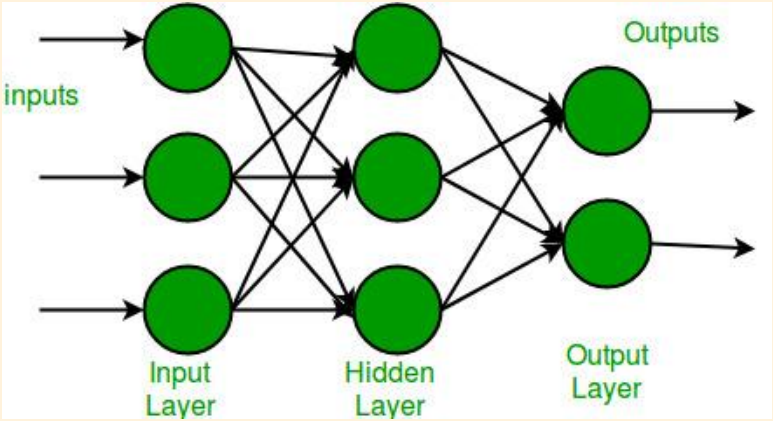
Result Re-Ranker

Why Re-Ranking

- Arabizi is highly ambiguous: one Romanized input can map to multiple Arabic forms
- Beam search often contains the right answer, but not always at rank 1
- Re-ranking uses extra candidate signals beyond raw sequence likelihood

Re-Ranking Model

- Lightweight MLP trained separately for each direction
- Uses beam candidates generated by the Transformer
- Training capped at 5,000 source rows per direction for compute efficiency
- Runs after beam search to reorder the top candidate list



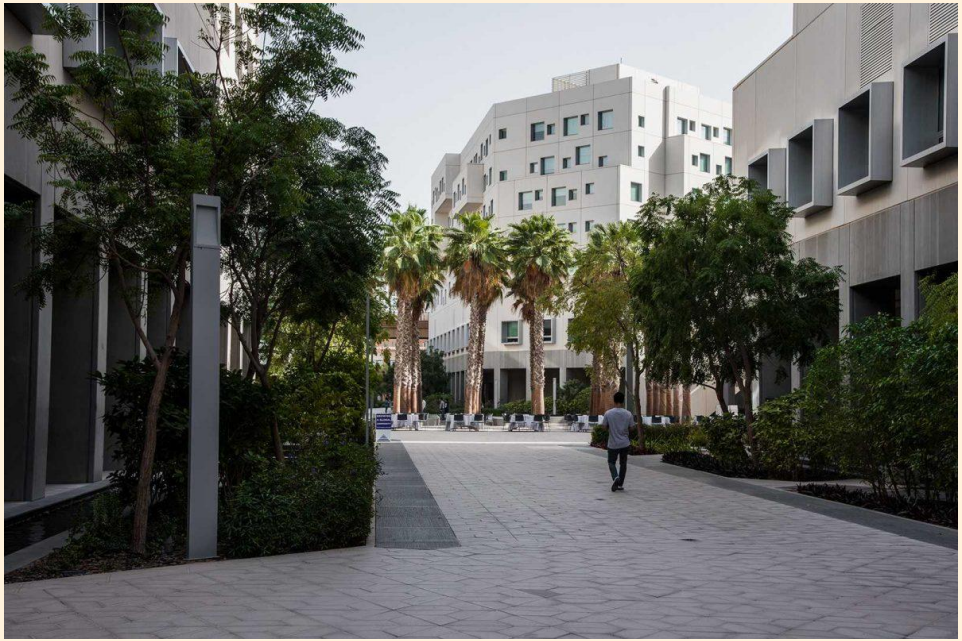
Candidate Features

- Beam rank, score gap, raw and normalized beam scores
- Source/candidate length features and ratios
- Arabizi digit usage: 2, 3, 5, 6, 7, 8, 9
- Space, vowel-like, and script-character patterns
- Cached lexicon features: 50k Arabic words + 26,740 Arabizi forms

Impact

- Improves user-facing top-k alternatives while keeping inference lightweight

Training and Data Preprocessing



Dataset: Maknuune v1.0.1

- From NYU Abu Dhabi
- Parallel Palestinian Arabizi-Arabic pairs
- 34,816 cleaned paired rows after deduplication/cleaning
- Pair split: 26,754 train / 3,344 validation / 3,345 test

Preprocessing Steps

- Unicode normalization (NFC form)
- Arabic tatweel/kashida removal
- Whitespace normalization
- Text stripping and cleaning
- Direction-specific tokenization
- Arabizi-aware augmentation for training variants

Task Construction

- Arabic→Arabizi train rows: 26,740
- Arabizi→Arabic train rows: 54,681 after augmentation

Model Performance

Metrics

- Normalized exact match and practical exact match
- Top-k beam hit rate and mean reciprocal rank (MRR)
- Direction-specific evaluation: Arabizi→Arabic and Arabic→Arabizi
- Full held-out test set: 6,686 direction-specific examples

Final Test Results

- Overall strict exact match: 69.91%
- Overall practical exact match: 72.18%
- Top-k strict exact match: 88.17%
- Top-k practical exact match: 89.64%
- Average edit distance: 1.28

By Direction

- Arabizi→Arabic: 72.36% strict; 76.91% practical; 90.88% top-k practical
- Arabic→Arabizi: 67.45% strict; 88.39% top-k strict

Beam Search & Re-Ranking

- Beam top-3 strict hit rate: 82.23%
- Beam top-5 strict hit rate: 85.13%
- Beam top-10 strict hit rate: 87.65%
- Beam MRR: 0.766

ع **Arabizi ↔ Arabic Translator**

Arabizi → Arabic

Palestinian Arabic

Arabizi Type Arabizi text

salaam

Clear | Copy Results | Transliterate

Dialect options

Palestinian Arabic Available		Levantine Arabic	Coming later
Egyptian Arabic	Coming later	Gulf Arabic	Coming later
Iraqi Arabic	Coming later	Maghrebi Arabic	Coming later

Arabic All returned transliterations · 0:08

1. سلام
2. صلام
3. لسام
4. ثلام
5. سلامة
6. سلامي

ع **Arabizi ↔ Arabic Translator**

Arabic → Arabizi

Palestinian Arabic

Arabic Type Arabic text

سلام

Clear | Copy Results | Transliterate

Dialect options

Palestinian Arabic Available	Levantine Arabic Coming later
Egyptian Arabic Coming later	Gulf Arabic Coming later
Iraqi Arabic Coming later	Maghrebi Arabic Coming later

Arabizi All returned transliterations · 0:07

1. salaam
2. salam
3. saalim
4. silmaam
5. s2aalam
6. sal2aam

ع **Arabizi ↔ Arabic Translator**

Arabizi → Arabic

Palestinian Arabic

Arabizi Type Arabizi text

keef

Arabic All returned transliterations · 0:07

1. كيف
2. كيفة
3. كافة

Clear | Copy Results | Transliterate

Dialect options

Palestinian Arabic Available	Levantine Arabic Coming later
Egyptian Arabic Coming later	Gulf Arabic Coming later
Iraqi Arabic Coming later	Maghrebi Arabic Coming later

ع **Arabizi ↔ Arabic Translator**

Arabic → Arabizi

Palestinian Arabic

Arabic Type Arabic text

كيف

Clear | Copy Results | Transliterate

Dialect options

Palestinian Arabic Available		Levantine Arabic Coming later	
Egyptian Arabic Coming later		Gulf Arabic Coming later	
Iraqi Arabic Coming later		Maghrebi Arabic Coming later	

Arabizi All returned transliterations · 0:07

1. kafi
2. kiif
3. fiik
4. keef
5. ykaf
6. kafa

ع **Arabizi ↔ Arabic Translator**

Arabizi → Arabic

Palestinian Arabic

Arabizi Type Arabizi text

qahwe

Clear | Copy Results | Transliterate

Dialect options

Palestinian Arabic Available		Levantine Arabic	Coming later
Egyptian Arabic	Coming later	Gulf Arabic	Coming later
Iraqi Arabic	Coming later	Maghrebi Arabic	Coming later

Arabic All returned transliterations · 0:09

1. قهوة
2. قحوة
3. حقوة
4. قهوةتة
5. قهوةزة

ع **Arabizi ↔ Arabic Translator**

Arabic → Arabizi

Palestinian Arabic

Arabic Type Arabic text

قهوة

Clear | Copy Results | Transliterate

Dialect options

Palestinian Arabic Available		Levantine Arabic	Coming later
Egyptian Arabic	Coming later	Gulf Arabic	Coming later
Iraqi Arabic	Coming later	Maghrebi Arabic	Coming later

Arabizi All returned transliterations · 0:07

1. qahwe
2. qahwwe
3. hwuqkshe
4. waqahwe
5. hwuqkkshe
6. qahwekahwe

goecknerwald@cua.edu

https://github.com/kagoeckner/arabizi_arabic_transliterator

https://kagoeckner.github.io/arabizi_arabic_transliterator/