Introduction

• Arabic is a collection of dialectal variants that are historically related but significantly different.
• These differences can be seen across regions, countries, and even cities in the same countries.
• Previous work on Arabic Dialect identification has focused mainly on specific dialect levels.
• We define a unified hierarchical schema for dialectal Arabic identification.

Arabic Linguistic Challenges

• Extensive orthographic, phonological, morphological and lexical variations among dialects.
• Arabic speakers tend to code-switch between their dialect and Modern Standard Arabic.

Unified Labeling of Arabic Dialect Data Sets

• Data Selection
  • Exclusively Arabic script data sets in Arabic dialects, with some degree of identification.

• Data Variability
  • Wide range of genres: speech transcripts, social media texts (tweets, news comments, youtube comments), SMS, forum novels, travel phrases, and song lyrics.
  • Labels vary widely in terms of granularity and spread.
  • Different sizes.

• Unified Labeling into a three level hierarchy: Region→Country→City

Results

• Aggregated Arabic dialect data from different sources into a unified schema
• Built aggregated character and word language models for dialect identification
• Models and mapping code are publicly available https://github.com/CAMeL-Lab/HierarchicalArabicDialectID

• Improved results on the city, country, and region levels by extending a SOTA technique for Arabic dialect id (Salameh et al., 2018).