Lexical Modeling of Egyptian Arabic for Automatic Speech Recognition

Taha Merghani[1] , Tuka Al Hanai[2] and James Glass[2]

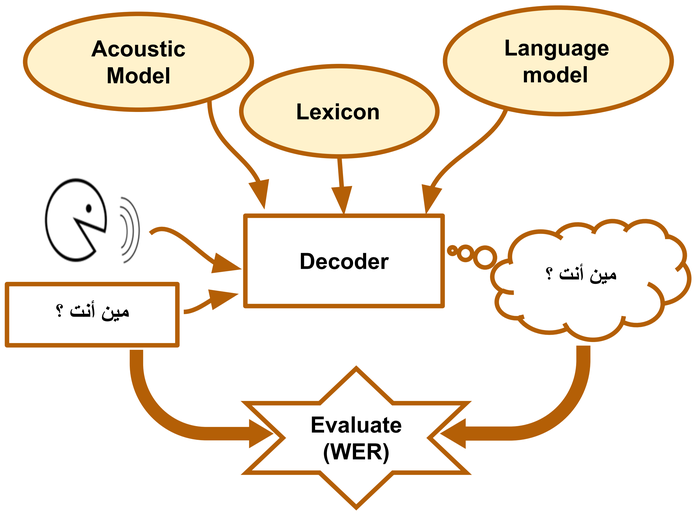
[1]Department of Electrical & Computer Engineering, Jackson State University

[2]MIT Computer Science & Artificial Intelligence Laboratory, Massachusetts Institute of Technology

1. Overview

- Speech-enabled applications growing in everyday life (Apple Siri, Car Navigation, Medical Transcription).
- 6000+ languages in the world.

Core of these systems is a recognizer:



Developing an Automatic Speech Recognition system (ASR) for Arabic is challenging.

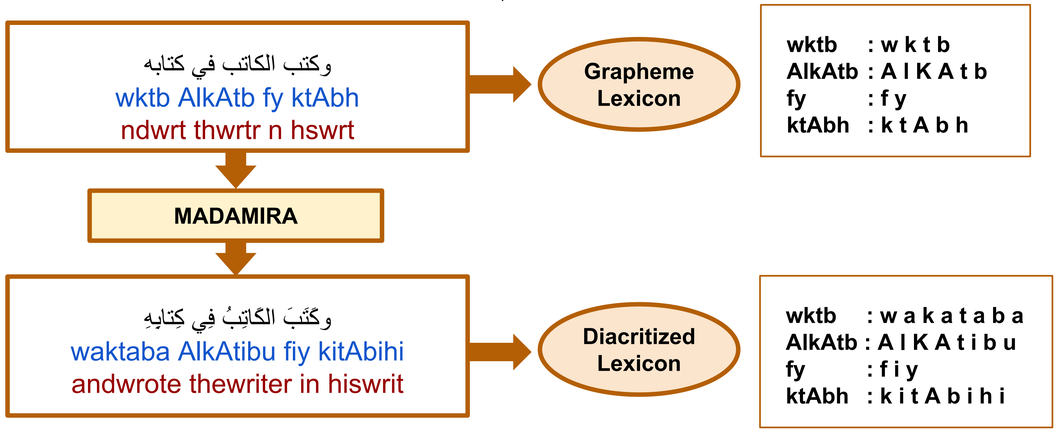
- **Ambiguous orthography**
  - No Diacritics --information is missing.
  - Ambiguous word to phoneme mapping.
- **Rich Morphology**
  - Large number of combinations of stems+affixes.
- **Diverse Dialects**
  - Arabic widely spoken, 250+ million speakers.
  - No standardized orthography.
  - Limited data.

2. MSRP Project

Research Question: Does diacritization of words improve ASR performance of Egyptian Arabic?

3. Methodology

- **Setup**
  - Corpus: Al-Jazeera Egyptian Dialect.
  - 81K Tokens, 18K Vocab.
  - Training set: 10 hrs, 1hr development set, and 1hr evaluation set.
  - Toolkits: Kaldi Speech Recognition, MADAMIRA text-processing.
  - Original transcription parsed and normalized to generate CODA transcription.
- **Lexical Modelling**
  - Diacritized lexicon with the aid of MADAMIRA, and Kaldi.



- **Language Modelling**
  - Trigram with SRILM Kneser-Ney discounting.
  - Built over training data transcript.
- **Acoustic Modeling**
  - Features: MFCCs + CMVN + LDA + MLLT + SAT.
  - Triphone context-dependant GMM-HMM.
  - Deep Neural Networks (DNN).
  - Sequence Deep Neural Networks (SDNN).

4. Results

Acoustic Model	Diacritized Lexicon		Grapheme Lexicon (CODA)		Grapheme Lexicon (Original)	
	Dev (%)	Eval (%)	Dev (%)	Eval (%)	Dev (%)	Eval (%)
GMM-HMM	56.8	59.1	57.4	58.6	60.0	61.5
DNN 1024 x 4	54.8	57.4	55.0	58.0	58.8	61.2
SDNN 1024 x 4	53.1	56.1	**53.0**	**55.6**	57.0	59.8
Lexicon Size	17.8K		17.5K		18.7K	
OOV (%)	15.3		15.3		16.6	
#Lexical Units	43		36		35	

5. Evaluation

Comparison between reference (REF) and hypothesis (HYP) generated by ASR.

Word Error Rate (WER) = (Subs + Dels + Ins) / N

- Subs: Number of substitutions between REF and HYP.
- Dels: Number of deletions in HYP compared to REF.
- Ins: Number of insertions in HYP compared to REF.
- N: Total number of words in REF.

6. Conclusion

- A graphemic lexicon (CODA) outperformed the diacritized lexicon.
- Reducing OOV improves WER.
- Using DNN and SDNN acoustic models improves WER.

7. Further Work

- **Train new acoustic models**
  - Long Short-Term Memory Recursive Neural Networks (LSTM-RNN) using Stacked Bottleneck features (SBN).
- **Morpheme Based Language Modeling**
  - Reduces Out-of-Vocabulary (OOV) rate of Dev/Eval sets.
- **Explore larger lexicon sizes using tweets**

References

- Pasha et al, “MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic”, 2014
- F. Biadsy, N. Habash, and J. Hirschberg, “Improving the Arabic pronunciation dictionary for phone and word recognition with linguistically-based pronunciation rules,”, 2009