Generating a Lexicon for the Hijazi Dialect in Arabic

Fatimah Alqahtani – Jazan University & RMIT University

Mark Sanderson - RMIT University

Outline

- Introduction
- The Basic Syntax of the Hijazi Dialect
- Experiments
- Result
- Conclusion

Introduction

- Kingdom of Saudi Dialect
  - Najdi , Hijazii, Gulf, Northern and Southern
- Hijazi Dialect
  - It is considered the dialect of western cities
  - The second place in terms of population.
  - Low-resource dialect
    - Obstructs NLP tool
    - Obstructs application of NLP (translation – identification)

In this research we generate a lexicon of Hijazi Dialect Manually and automatically

The Basic Syntax of the Hijazi Dialect

- Irregular
  - Lexical.
    - Ex: Good, كويس “kwys” in Hijazi, طيب “Tyb” in MSA
  - Independent Pronouns.
    - Ex: 'AHnA' احنا or 'nHnA' نحنا in Hijazi, but it is 'nHn' نحن in MSA
- Regular
  - Negation.
    - The 'mA’ ما prefix with verbs and the word ‘mw’مو with adjectives and nouns,
  - Verb

Tens	MSA	Hijazi
Present	يقف ‘yqf’ he stand up’.	بيوقف ‘bywqf’
Future	يقول راح ‘rAH yqwl’ he’ll say / سو ف ‘swf’ ‘will’	حيقول ‘Hyqwl’

  - Letter Substitution.

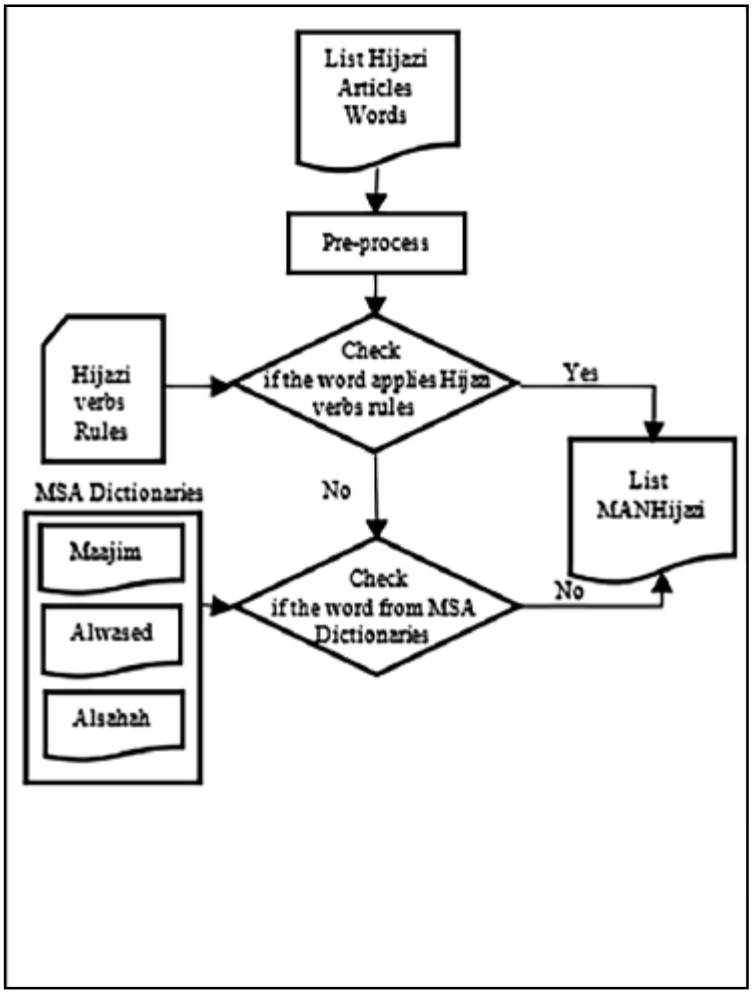
Letter in MSA	Letter substitution in Hijazi	Example in MSA	Example in Hijazi
‘ذ’ ‘*’	‘ز’ z / ‘د’ d	‘*kryAt’ (memories)/ ‘k*Ab’ (liar)	‘zkryAt’/ ‘kdab’
‘ث’ ‘v’	‘ت’ t / ‘س’ s	‘vqyl’ (heavy)/ ‘mvAl’ (example)	‘tqyl’/ ‘msAl’

Experiments

- Manual
  - Articles, Tweets
  - 3 Arabic dictionaries.
  - Hijazi Morphological
- Automatic techniques
  - Adapted an approach by Darwish et al. (2014),
  - Two roots sets (Sebawi, Quranic)
  - Hijazi Morphological.

Experiments

- Manual
  - Articles.
    - 156 Hijazi articles(Feb. 2011 to Sep. 2014).
    - The set contained 59,225 tokens.
    - Pre-process: remove MSA(Ranks NL)
    - HIJD morphological.
    - 3 MSA Dictionaries(Maajim, Alwased, Alsahah)
  - Result
    - 1,363 MANHijazi words.

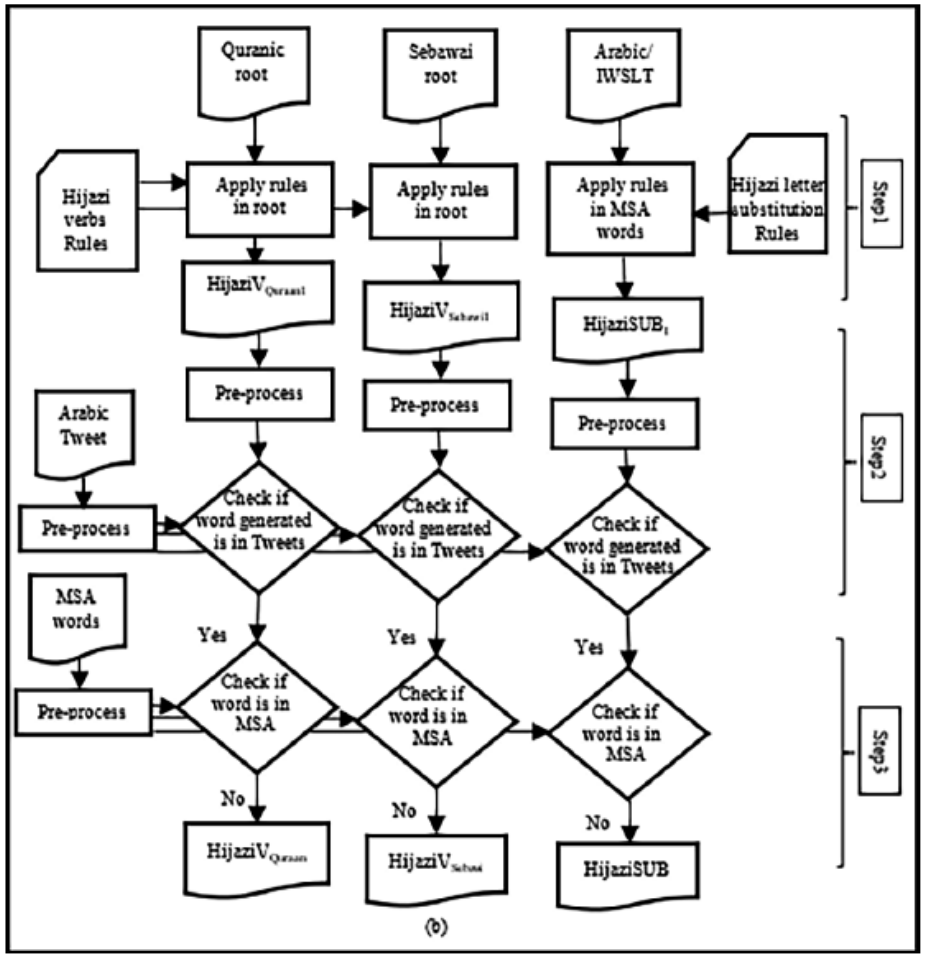


Experiments

- Manual
  - Tweets.
    - 3000 tweets have geo-location west of SA from Mourad et al.(2017)
    - Annotated by 3 native speaker.
    - Label as HIJD and non HIJD, determine the HIJD words
    - In one week with offer $50.
  - Result
    - We used Fleiss` Kappa to calculate the agreement 0.89.
    - 372 Hijazi Dialect tweets.
    - 666 HIJD words in tweets, 305 unique HIJD words

Experiments

- Automatic
  - Designed to be generalize to build and extend the data
  - To generate 3 lexicon lists of Low Resource Hijazi Dialect
  - It has 3 steps:
    - Step 1: Automatic generating by applying rules
    - Step 2: Filtering by tweets tokens
    - Step 3: Filtering by MSA words
  - Result
    - $HijaziV_{Sebawai:}$ 24,413
    - $HijaziV_{Quraan:}$ 12,428
    - $Hijazi_{SUB}$: 1,074



Results

- Evaluation:
  - We manually investigated the correcting of a 2% sample from $HijaziV_{Sebawai}$, $HijaziV_{Quraan}$, and $Hijazi_{SUB}$
  - We analysed the error rate
    - Alternative root: the word اتعمش 'AtEm$', 'weak eyesight'→ اتعمى 'AtEmY'
    - Incorrect root: ملم 'mlm', 'has knowledge', the root is a noun according to the dictionary not a verb root
    - Error a rule: الرضا 'AlrDa' 'satisfaction', ض 'D' changed to ز 'z' to become 'Alrza'

Type of list	**No. words**	**Sample words 2%**	**Hijazi words in 2%**	**Non-Hijazi words in 2%**	**Error rate**	**Distribution error rate**		
						Alternative Hijazi root	In-correct root	Error rule
$HijaziV_{Sebawai}$	**24,413**	490	409	81	0.16	69	6	6
$HijaziV_{Quraan}$	12,428	250	227	23	**0.09**	14	0	8
HijaziSUB	1,074	22	16	6	0.27	0	0	6

Conclusion

- The Morphological rules in the Hijazi are the first step to build a Hijazi lexicon.
- We can apply an automatically approach from High-Resource to Low-Resource to generate a lexicon (3 lexicon lists)
- Our approach more generalizes for Low-resource Semantic Languages/Dialect.


Questions