“Towards Building a Corpus for Palestinian Dialect”

Diyam Akra
Mustafa Jarrar
April, 2014
Outline

• Introduction
• Literature Review
• Research Methodology and progress
• Conclusions and next step
Introduction
Scope and Motivation

• Processing of written Palestinian Dialect
• Standard Arabic have rules but Dialects have no rules (differ in morphological, lexical, orthographical, phonological)

• Examples :

"سكر الباب", "هاظا جعفر" , "حاسها مش طايقتني", "اصلا بلبقلوش"

• Importance of building annotated corpus: Information Retrieval and extraction, search engines, translation, Auto Complete, Part_Of_Speech, parsers

• Annotation: annotate word with relevant meta data; here in our research we annotate it with the rules that applies and morphological information.
Introduction
Problem statement and thesis goals

• Problem statement:

*collect and annotate a corpus of Palestinian Dialect with relevant Meta data.*

• specific goals:
1. Collect written text of Palestinian Dialect.
2. Develop tools to automatically annotate the corpus.
3. Annotate the corpus manually.
4. Evaluate the automatic annotation with manual annotations
Summary of contributions

• Until to this moment, our contributions:

1. Collecting written text of Palestinian Dialect from different resources.

2. Parsing and indexing the written text automatically and storing it in a proper schema.

3. Develop some modules of automatic tools for processing and cleaning the collected corpus.
Introduction

Summary of contributions

• In the next stage:

1. Annotate the corpus manually.

2. Develop tools to automatically detect dialect words and phrases and annotate the corpus with Meta data.

3. Evaluating the automatic annotation (i.e., our tools) by comparing it with the manual annotations; to measure the accuracy of our automatic annotator
Literature Review

• MAGEAD

• MAGEAD is a morphological analyzer for Arabic and its Dialects

• MAGEAD idea: Build Morphological Behavior classes for every lexeme in Arabic and its Dialects

• Example: ٍازدَهَرت → "<zhr, AV1tV2V3, iaa> + at"

• published on (Habash, Rambow, Kiraz/2005; Habash, Rambow /2006; Altantawy, Habash, Rambow, Saleh /2010)
Literature Review

• Parsing Arabic Dialects

• Levantine Dialect Arabic Treebank, small Levantine_MSA dictionary, MSA tagger

• Three approaches for parsing: sentence transduction, Treebank transduction, grammar transduction

• Old work, stop go in this approach

• published on (Habash, Rambow, Kiraz/2005; Habash, Rambow /2006; Altantawy, Habash, Rambow, Saleh /2010)
Literature Review

- COLOBA project

- COLOBA: dialect data as input $\rightarrow$ retrieve all relevant information over web

- COLOBA idea: use many tools such as DI Identification Pipeline, COLANN_GUI which is a web application for annotate dialect data, DIRA, MAGEAD.

- Example: أصبحوا-يصبحوا-بيصبحوا-بقى-بقى$\rightarrow$ أصبح

- published on (Habash, Rambow, Diab, Kanjawi-Faraj/2008; Benajiba, Diab/2010; Diab, et-al./2010)
• CODA: Conventional Orthographic for Dialectal Arabic

• CODA is a standard for writing dialect Text

• CODA idea: dialect text as input $\rightarrow$ return in CODA standard

• Example: $\text{بردة} \rightarrow \text{برضه}$

• published on (Habash, Diab, Rambow/2012)
Literature Review

• CALIMA is a Morphological Analyzer for Egyptian Arabic

• CALIMA idea: building on the top of Egyptian Colloquial Arabic Lexicon (ECAL)

• CALIMA has six tables: (complex prefix, complex suffix, stem, prefix-stem, prefix-suffix, stem-suffix)

• CALIMA has: complex-prefix entries is 2421, complex-suffix entries 1179, stem entries 100000

• published on (Habash, Eskander, Hawwari/2012)
Research Methodology and progress

1. **Collect Palestinian Dialect Data**: different resources which are: books, Palestinian series “وطن ع وتتر” , social networks; all of it manually collected (done)

2. **Collect Palestinian Dialect Patterns**: classify rules presented in “معجم العامي و الدخيل” then rewrite it in algorithmic way. (done)

3. **Design a database for storing the corpus data**: store data using N-gram model (1-4), store position in order to regenerate (done)
4. Build a gold Standard Dataset:

a) **exclude all stop words**, less than three characters words, words with special characters and non Arabic **(almost done)**

b) **Filter with Standard Arabic** **(almost done)**

c) **Generate stem, prefix, and suffix** using Khoja stemmer; in order to detect all possible pre, in, suf-fixes to detect new patterns **(done)**
Research Methodology and progress

5. **Automatic annotation**: design an algorithm to detect dialect patterns in step 2; annotate word with pattern that applies. *(not done)*

6. **Manual annotation**: manually detect and annotate dialect terms/phrases. *(not done)*

7. **Evaluation**: words that correctly classified according to manual annotation, words that wrongly classified according to manual annotation. *(not done)*
## Corpus statistics

<table>
<thead>
<tr>
<th>Document Type</th>
<th>Number of pages/ or words</th>
</tr>
</thead>
<tbody>
<tr>
<td>شهود النكبة“ book</td>
<td>114 pages</td>
</tr>
<tr>
<td>Facebook</td>
<td>3947 words</td>
</tr>
<tr>
<td>Twitter</td>
<td>3812 words</td>
</tr>
<tr>
<td>Blogs</td>
<td>9087 words</td>
</tr>
<tr>
<td>Palestinian Stories</td>
<td>2772 words</td>
</tr>
</tbody>
</table>
## Corpus statistics

<table>
<thead>
<tr>
<th>Document Name</th>
<th>Number of pages/words</th>
</tr>
</thead>
<tbody>
<tr>
<td>PalDF <a href="http://www.paldf.net">www.paldf.net</a></td>
<td>2123 words</td>
</tr>
<tr>
<td>&quot;وطن ع وتر&quot; Series</td>
<td>23057 words</td>
</tr>
<tr>
<td>Dictionary of Palestinian Vocabularies and Loan words</td>
<td>646 pages / 5595 words</td>
</tr>
</tbody>
</table>
## Final Results

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Number of Distinct words</td>
<td>15372</td>
</tr>
<tr>
<td>Total Number of Distinct words that are in our corpus and also found in Ramooz</td>
<td>3187</td>
</tr>
<tr>
<td>Total Number of Distinct words that are in our corpus and not found in Ramooz; here we assume it totally Dialect</td>
<td>12185</td>
</tr>
<tr>
<td>Total Number of Distinct stem</td>
<td>4276</td>
</tr>
<tr>
<td>Total Number of Distinct prefix</td>
<td>274</td>
</tr>
<tr>
<td>Total Number of Distinct suffix</td>
<td>391</td>
</tr>
</tbody>
</table>
Conclusions

• irregular prefixes or suffixes, they are either dialect words or not and maybe it used to extract new patterns for Palestinian dialect

• list of dialect words still have MSA words according to list of prefixes or suffixes
Next Step

• improve unique Dialect list according to prefixes, and suffixes list

• Developing tools to automatically annotate corpus.

• build a small application that would take Palestinian Text as Input; return it in Standard Arabic as Output

• analyze the results
دعوة

 حدورنا الكريم: إذا كان لديك أي نصوص بالعامية الفلسطينية وتودون مشاركتها نكون شاكرين لكم.
شكر

نود أن نشكر كل من:

الدكتور نزار حبش

الدكتور مهدي عرار

على مساهماتهم في المشروع حتى الآن

* كما نود أن ننوه أن هذه المشروع بالشراكة مع كل من الدكتور نزار حبش و

الدكتور مهدي عرار
References


References


References


References


