جامعة بيرزيت
BIRZEIT UNIVERSITY

Birzeit University

“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(بيلعب-يلعب), lexical(بس-بدي) , orthographical(م بقلكش-ما بقلكش) , phonological(ق-أ-ك-غ))

- 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

Introduction
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 : **ازدهرت→"<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 → 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 : أصبح→ أصبحوا ـ يصبحوا-حيصبحوا-بيصبحوا-بقى-بيبقى

- published on (Habash, Rambow, Diab, Kanjawi-Faraj/2008; Benajiba, Diab/2010;Diab, et-al./2010)

Literature Review

- CODA : Conventional Orthographic for Dialectal Arabic
- CODA is a standard for writing dialect Text
- CODA idea : dialect text as input→ return in CODA standard
- Example : بردة→ برضه, يعجبوكوا→ يعجبوكو
- 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 وتر” , 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)

Research Methodology and progress

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

Document Type	Number of pages/ or words
“شهود النكبة” book	114 pages
Facebook	3947 words
Twitter	3812 words
Blogs	9087 words
Palestinian Stories	2772 words

Corpus statistics

Document Name	Number of pages/words
PalDF www.paldf.net	2123 words
“وطن ع وتر” Series	23057 words
Dictionary of Palestinian Vocabularies and Loan words	646 pages / 5595 words

Final Results

Total Number of Distinct words	15372
Total Number of Distinct words that are in our corpus and also found in Ramooz	3187
Total Number of Distinct words that are in our corpus and not found in Ramooz; here we assume it totally Dialect	12185
Total Number of Distinct stem	4276
Total Number of Distinct prefix	274
Total Number of Distinct suffix	391

Conclusions

- irregular prefixes or suffixes, they are either a 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

* حضورنا الكريم : إذا كان لديكم أي نصوص بالعامية الفلسطينية و تودون 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

[1] Habash, N., Rambow, O., and Kiraz, G. 2005.Morohological analysis and generation for Arabic Dialects. Semitic '05 Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages. Pages 17-24

[2] Habash, N., and Rambow, O. 2006.MAGEAD: Morphological Analyzer and Generator for Arabic Dialects. In proceedings of the 21st International conference on Computational linguistics and 44th Annual Meeting of ACL, pages 681-688, Sydney, July 2006.

[3] Altantawy, M., Habash, N., Rambow, O., and Saleh, I. 2010. Morphological Analysis and Generation of Arabic Nouns: A Morphemic Functional Approach. In Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta; 01/2010.

[4] Almeman, K., and Lee, M. 2012.Towards Developing a Multi-Dialect Morphological Analyzer for Arabic. In Proceedings of 4th International Conference on Arabic Language Processing, May 2-3, 2012, Rabat, Morocco.

[5] Rambow, O., Chiang, D., Diab, M., Habash, N., Hwa, R., Sima’an, K., Lacey, V., Levy, R., Nichols, C., and Shareef, S. 2006.Parsing Arabic Dialects. Final Report-Version 1,January 18,2006.

References

[6] Maamouri, M., Bies, A., Buckwalter, T., Diab, M., Habash, N., Rambow, O., and Tabessi, D. 2006. Developing and using a Pilot Dialectal Arabic Treebank. In Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC’06

[7] Chiang, D., Diab, M., Habash, N., Rambow, O., and Shareef, S. 2006.Parsing Arabic Dialects . In Proceedings of European Chapter of the Association for Computational Linguistics(EACL), Trento, pages 369–376.

[8] Habash, N., Rambow, O., Diab, M., and Kanjawi-Faraj, R. 2008.Guidelines for Annotation of Arabic Dialectness. In Proceedings of workshop on Arabic and its local Languages,Marrakech,Mococco,2008.

[9] Benajiba, Y., and Diab, M. 2010.A Web Application for Dialectal Arabic Text Annotation. In Proceedings of the LREC Workshop for Language Resources (LRs) and Human Language Technologies (HLT) for Semitic Languages: Status, Updates, and Prospects. 2010

References

[10] Diab, M., Habash, N., Rambow, O., Altantawy, M., and Benajiba, Y. 2010. COLABA: Arabic dialect annotation and processing. In Proceedings of LREC Workshop on Semitic Language Processing. 2010. p. 66-74

[11] Graja, M., Jaoua, M., and Belguith, L.2011. Building Ontologies to Understand Spoken Tunisian Dialect. In Proceedings International Journal of Computer Science, Engineering and Application(IJCSEA) Vol.1,No.4,August 2011.

[12] Diab, M., and Habash, N.2009. Presentation about Arabic Dialect Processing. MEDAR 2009, Cairo, Egypt, April 21, 2009

[13] Khoja, S. Using Khoja Stemmer. Available at http://zeus.cs.pacificu.edu/shereen/research.htm

[14] Lubany, A. Dictionary of Palestinian Vocabularies and Loan Words (Arabic-Arabic) عربي–معجم العامي و الدخيل في فلسطين عربي.2006.Lebanon Library.

References

[16] Habash, N., Eskander, R., and Hawwari, A. 2012.A Morphological Analyzer for Egyptian Arabic . In Proceedings of Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology (SIGMORPHON2012) . 2012. p. 1-9

[17] Habash, N., Diab, M., and Rambow, W.2012. Conventional Orthography for Dialectal Arabic . In Proceedings ..Vol.1,No.4,August 2011**.**

[18] Habash, N., Roth, R., Rambow, W. , Eskander, R., and Tomeh, N. 2013.A Morphological Analysis and Disambiguation for Dialectal Arabic . In Proceedings of NAACL-HLT. 2013. p. 426-432

[19] Eskander, R., Habash, N., and Rambow, W. Automatic extraction of Morphological lexicons from Morphologically Annotated Corpora . In Proceedings 2013 Conference on Empirical Methods in Natural Language Processing . 2013. p. 1032-1043