Parsing Arabic Dialects

Owen Rambow
Columbia University
rambow@cs.columbia.edu
Team

• Senior Members
  o David Chiang U of Maryland
  o Mona Diab Columbia
  o Nizar Habash Columbia
  o Rebecca Hwa U of Pittsburgh
  o Owen Rambow Columbia
  o Khalil Sima'an U of Amsterdam

• Grad Students
  o Roger Levy Stanford
  o Carol Nichols Pittsburgh
Team (ctd)

- Undergrads
  - Vincent Lacey, Georgia Tech
  - Safiullah Shareef, Johns Hopkins

- Externals
  - Srinivas Bangalore, AT&T Labs -- Research
  - Martin Jansche, Columbia
  - Stuart Shieber, Harvard
  - Richard Sproat, U of Illinois at UC
  - Bill Young, CASL/U of Maryland
Overview

• Team
• **Problem: Why Parse Arabic Dialects?**
• Proposed Approaches
• Accomplishments to Date
• Plan for Workshop
• Summary
didn't buy Nizar new table

Nizar bought new table

Nizar bought new table

Nizar didn't buy new table
Factors Affecting Dialect Usage

- Geography (continuum)
- City vs village
- Bedouin vs sedentary
- Religion, gender, ...

⇒ Multidimensional continuum of dialects
Lexical Variation

| English | table | cat | of | (I) want | there is | there isn’t |
|---------|-------|-----|----|----------|----------|-------------|
| MSA     | Tāwila | qiTTa | idafa | ‘uridu | yujadu | la yujadu |
| Moroccan| mida  | qeTTa | dyal | bgit     | ḳayn    | ma kāynš  |
| Egyptian| Tarabīza | ‘oTTa | bita3 | 3awez | fi     | ma fi |
| Syrian  | Tawle | bissē | taba3 | biddi | fi     | mā fi |
| Iraqi   | mēz   | bazzūna | mal | ‘arid  | aku     | māku       |
And you didn’t write it for him
Dialect Syntax: Word Order

- **Verb Subject Object**
  
  كتاب الأولاد الأشعار
  
  wrote.masc the-boys the-poems (MSA)

- **Subject Verb Object**
  
  الأولاد كتب الأشعار
  
  the-boys wrote.masc.pl the-poems (LEV, EGY)

|          | Most common word order | Full agreement in VSO | Full agreement in SVO |
|----------|------------------------|------------------------|------------------------|
| MSA      | VSO                    | no                     | ✓                      |
| Dialects | SVO                    | ✓                      | ✓                      |
Dialect Syntax: Noun Phrases

• Possessives
  o Idafa construction
    ➢ $\textbf{Noun1 Noun2}$
    ➢ ملك الأردن
      king Jordan
      the king of Jordan / Jordan’s king
  o Dialects have an additional common construct
    ➢ $\textbf{Noun1 <particle> Noun2}$
    ➢ LEV: الملك تبع الأردن
      the-king belonging-to Jordan
    ➢ <particle> differs widely among dialects

• Pre/post-modifying demonstrative article
  o MSA: هذا الرجل
    this the-man
    $this\ man$
  o EGY: الراجل ده
    the-man this
    $this\ man$
Code Switching:
Al-Jazeera Talk Show

MSA and Dialect mixing in formal spoken situations

لا أنَّ ما يعتقد لأنه عملية اللي عم يعارضوا اليوم تمديد للرئيسي لجودهم اللي طالبوا بالتمديد للرئيسي والنروي وبالتالي موضوع منه موضوع متدني على الأرض أنا يحترم أنه يكون في نظرة الديمقراطية للأمور وأن يكون في احترام لعبة الديمقراطية وأن يكون في ممارسة ديمقراطية وبيند يتجذب عليه أيضاً السماح في لبنان لرئيسي هذا الموضوع، ليس بدي برجع لحظة على موضوع إنجازات العهد يعني نعم نحكي عن إنجازات السادات لكن هل النظام في لبنان نظام رئيسي النظام في لبنان من بعد الطائف ليس نظام رئيسي وبالتالي السلطة هي عملياً بيد الحكومة مجتمعة والرئيسي لجود أثبت خلال ممارسته الأخيرة بأنه لما يكون في شخص مسؤول في منصب معين وأن عشت هذا الموضوع شخصياً بممارساتي في موضوع الاتصالات لما بأخذ موقف صاحب ضمن خطاب ومبادئ خطاب القسم هو إلى جانب أنه مش مطلوب من رئيس جمهورية هو يكون رئيس السلطة التنفيذية لأنه منه في لبنان ما بعد إتفاق الطائف رئيس السلطة التنفيذية عليه التوجيه عليه إبداء الملاحظات عليه القول ما هو خطأ وما هو صريح عليه تثبيت وجهة الوطن الشاملة كي يظل في مسؤولية وطنية كي يظل في نفسي ما بين المسلم والمسيحي في لبنان يحتضن أنباء هذا البلد ما يترك المسار يروح باتجاه الخطأ نعم إما خطاب القسم كان موضوع مبادئ طرحت هو ملزم فيها اللي يمشوا معه وأمموا فيها الزموا فيها أنا أثبت خلال الأربعة سنوات بالแนะنيدة الحكومية أني التزمت فيها وما التزمنا بهذا الموضوع كان الرئيس لجود إلى جنبنا في هذا الموضوع، أما الموضوع الديموقراطي أنا نفهم تماماً هذا هالوجهة النظر ليس ما يمكن نقول إنه الدستور أو تعديله هو أو إمكانية فتح إعادة انتخاب ديموقراطي ضمن المجلس والتصويت إلى ما هناك لرئيسي جمهورية بولاية ثانية هو مسح هيئة في جوهر الديمقراطية هذا بالاقل يعني قناعي في هذا الموضوع.
Why Study Arabic Dialects?

- There are no native speakers of MSA
- **Almost no** native speakers of Arabic are able to sustain continuous spontaneous production of spoken MSA
- This affects **all** spoken genres which are not fully scripted: conversational telephone, talk shows, interviews, etc.
- Dialects also in use in new media (newsgroups, blogs, etc)
Arabic Dialects: Computational Resources (1)

- Transcribed speech/transcript corpora
  - **Levantine** (LDC), Egyptian (LDC), Iraqi, Gulf, ...
- Very little other unannotated text
  - Online: Blogs, newsgroups
  - Paper: Novels, plays, soap opera scripts, ...
- Teensy treebanks
  - **Levantine**, LDC for this workshop with no funding
  - JUST FOR EVALUATION
Arabic Dialects: Computational Resources (2)

- Lexicons
  - CallHome Egyptian Arabic monolingual lexicon (LDC)
  - Egyptian-MSA and Levantine-MSA dictionary (Columbia, NSF funding)
  - Wordlists (CASL, web, published sources)

- Morphological resources
  - CallHome Egyptian Verb transducer (LDC)
  - Columbia University Arabic Dialect Project: MAGEAD: Pan-Arab Morphology, only MSA so far (ACL workshop 2005)
  - Buckwalter morphological analyzer for Levantine (LDC, under development)

- Contrast to MSA: huge corpora, treebank, lexicons, morphological analyzers, taggers, chunkers, parsers, MT system, ASR systems, ...
Overview

- Team
- Problem: Why Parse Arabic Dialects?
- **Proposed Approaches**
- Accomplishments to Date
- Plan for Workshop
- Summary
Parsing Arabic Dialects: Problem

- Dialect
  - بالاولاد كتيو الاشعار
  - كتيو
  - والاولاد الاشعار

- MSA
  - Treebank
  - Parser
  - Big LM
  - Small LM
 Parsing Arabic Dialects

- Many different dialects
- Dialects are spoken, few written resources
- Code switching
- Conclusion: Can’t assume we will get treebanks or even large unnatoated corpora for each dialect
- What to do?
Parsing Arabic Dialects

• Idea: use rich resources for MSA, adapt for dialects

• Exploit:
  o Similarity MSA-dialects
  o Explicit knowledge we have about differences

• Three approaches
Proposed Solution 1: Dialect Sentence Transduction
Proposed Solution 2: MSA Treebank Transduction

Dialect

Small LM

Treebank

Parser

MSA

Treebank

Parser

Big LM
Proposed Solution 3: MSA Grammar Transduction
Proposed Solution 3: MSA Grammar Transduction

Dialect

| Small LM |
|-----------|
| الاولاد كتب الاعمار |
| كتب الاعمار |
| الاولاد |

Probabilistic TAG

Parser

MSA

| Treebank |
|-----------|
| TAG = Tree Adjoining Grammar |

| Parser |
|--------|
| Big LM |
Overview

- Team
- Problem: Why Parse Arabic Dialects?
- Proposed Approaches
- Accomplishments to Date
- Plan for Workshop
- Summary
Accomplishments so Far:
All Three Approaches

- Baselines: using MSA parsers on Levantine (done/ongoing)
- Seed MSA/Levantine dictionary (by hand, ongoing)
- Finding Levantine text on the web (ongoing)
- Estimating language models for MSA and Levantine (done/ongoing)
- Estimating lexical translation probabilities without a parallel corpus (detailed proposal)
Accomplishments so Far:
Dialect-to-MSA Sentence Transduction

- Sentence transducer (implemented)
- Adapt sentence transducer to Levantine (ongoing)
- Backmapper for trees MSA-to-dialect (detailed proposal)
Accomplishments so Far: MSA-to-Dialect Treebank/Grammar Transduction

• Grammar extracted from MSA treebank (more work needed)
• Tree transformation rule formalism and rules for MSA-to-Levantine (implemented, under revision)
• Synchronous grammar (with explicit or hidden grammar): parallel parsing without a parallel corpus (detailed proposal)
• Treebank transfer: estimating parameters without a parallel corpus (detailed proposal)
Work Plan

• Implement all three proposals by end of week 2
• Determine how to proceed
Summary

- Continuum of dialects
- People communicate spontaneously in Arabic dialects, not in MSA
- So far no computational work on dialects, almost no resources (not even much unannotated text)
- Do not want ad-hoc solution for each dialect
- Exploit knowledge of differences MSA/dialects
