Parsing Arabic Dialects

JHU Summer Workshop
Final Presentation
August 17, 2005
Global Overview

- Introduction (Owen Rambow)
- Student Presentation: Safi Shareef
- Student Presentation: Vincent Lacey
- Lexicon
- Part-of-Speech Tagging
- Parsing
  - Introduction and Baselines
  - Sentence Transduction
  - Treebank Transduction
  - Grammar Transduction
- Conclusion
Team

- **Senior Members**
  - David Chiang  U of Maryland
  - Mona Diab      Columbia
  - Nizar Habash  Columbia
  - Rebecca Hwa  U of Pittsburgh
  - Owen Rambow  Columbia  (team leader)
  - Khalil Sima'an  U of Amsterdam

- **Grad Students**
  - Roger Levy  Stanford
  - Carol Nichols  U of 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
  - Otakar Smrz, Charles U, Prague
  - Richard Sproat, U of Illinois at UC
  - Bill Young, CASL/U of Maryland

Contact: Owen Rambow, rambow@cs.columbia.edu
Local Overview: Introduction

- Team
- **Problem: Why Parse Arabic Dialects?**
- Methodology
- Data Preparation
- Preview of Remainder of Presentation:
  - Lexicon
  - Part-of-Speech Tagging
  - Parsing
The Arabic Language

- Written language: Modern Standard Arabic (MSA)
- MSA also spoken in scripted contexts (news broadcasts, speeches)
- Spoken language: dialects
didn’t buy Nizar table new

Nizar məstarəfat tərabəza gididə

Nizar məstarəfat tawile ədəde

Nizar məfrəf mida ədəde

Nizar not-bought-not table new
Factors Affecting Dialect Usage

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

⇒ Multidimensional continuum of dialects
Lexical Variation

- Arabic Dialects vary widely lexically

<table>
<thead>
<tr>
<th>English</th>
<th>table</th>
<th>cat</th>
<th>of</th>
<th>(I)_want</th>
<th>there is</th>
<th>there isn`t</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSA</td>
<td>Tawila</td>
<td>qiTTa</td>
<td>idafa</td>
<td>`uridu</td>
<td>yujadu</td>
<td>la yujadu</td>
</tr>
<tr>
<td>Moroccan</td>
<td>mida</td>
<td>qeTTa</td>
<td>dyal</td>
<td>bgit</td>
<td>k`ayn</td>
<td>ma k`aynš</td>
</tr>
<tr>
<td>Egyptian</td>
<td>Tarabelza</td>
<td>`oTTa</td>
<td>bita3</td>
<td>3aweze</td>
<td>fi</td>
<td>ma fi</td>
</tr>
<tr>
<td>Syrian</td>
<td>Tawle</td>
<td>bisse</td>
<td>māl</td>
<td>biddi</td>
<td>fi</td>
<td>mafisš</td>
</tr>
<tr>
<td>Iraqi</td>
<td>mēz</td>
<td>bazzūna</td>
<td>`arid</td>
<td>aku</td>
<td>maku</td>
<td></td>
</tr>
</tbody>
</table>
Morphological Variation
Verb Morphology

And you didn’t write it for him
Dialect Syntax: Word Order

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

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

<table>
<thead>
<tr>
<th></th>
<th>VS Order</th>
<th>V Order</th>
<th>SV Order</th>
<th>Full agreement in VSO</th>
<th>Full agreement in SVO</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSA</td>
<td>35%</td>
<td>30%</td>
<td>35%</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Dialects</td>
<td>11%</td>
<td>62%</td>
<td>27%</td>
<td>yes</td>
<td>yes</td>
</tr>
</tbody>
</table>
Dialect Syntax: Noun Phrases

■ Possessives
  ■ Idafa construction  
    ■ Noun1 Noun2  
    ■ ملك الأردن
      king Jordan
      the king of Jordan / Jordan’s king
  ■ Dialects have an additional common construct  
    ■ Noun1 <particle> Noun2  
    ■ LEV: الملك يتبع الأردن: the-king belonging-to Jordan  
    ■ <particle> differs widely among dialects

■ Pre/post-modifying demonstrative article
  ■ MSA: هذا الرجل
    this the-man
  ■ EGY: الرجل ده
    the-man this
لا أنا ما بعتقد لأنه عملية اللي عم بيعرضوا اليوم تمديد للرئيس لحود هم
اللي طالبوا بالتمديد للرئيس الهراوي وبالتالي موضوع منه موضوع مبدئي على الأرض أنا بحترم أنه يكون في نظرة
ديمقراطية للأمور وأنه يكون في احترام للعبة الديمقراطية وأن يكون في ممارسة ديمقراطية ويعتقد إنه الكل في
لبنان أو أكثرية ساحقة في لبنان تردد هذا الموضوع، بينشي بدي يرجع لحجة على موضوع إنجازات العهد يعني نعم
نحكي عن إنجازات العهد لكن هل النظام في لبنان نظام رئاسي النظام في لبنان من بعد الطائف ليس نظام رئاسي
وبالتالي السلطة هي عملياً بيد الحكومة مجتمعة والنالدات لحود أيت خلال ممارسته الأخيرة بأنه لما يكون في
شخص مسؤول في منصب مهين وانا أعيش هذا الموضوع شخصياً بممارساتي في موضوع الاتصالات لما يأخذ
مواقف صالحة ضمن خطاب ومبادئ خطاب القسم هو إلى جانب حا مش مطلوب من رئيس جمهورية هو يكون
رئيس السلطة التنفيذية لأنه منه بقي في لبنان ما بعد إتفاق سلطان رئيس السلطة التنفيذية عليه التوجيه عليه إبداء
الملاحظات عليه القول ما هو خطأ وما هو صحيح عليه تنفيذ جهود الوطنية الشاملة كي يظل في مصافحة وطنية كي
يظل في تواق ما بين المسلم والمسيحي في لبنان يحتضن أبناء هذا البلد ما يترك المسار بروح باتجاه الخطأ نعم
إذا خطاب القسم كان موضوع مبادئ طرحت هو ملتزم فيها اللي مشي المعاً معه وأمنوا فيها التزموا فيها أنا أثبت
خلال الأربع سنوات بالمشاركة الحكومية أن السلم فيها واما التزمنا بهذا الموضوع كان رئيس لحود إلى جنبا
في هذا الموضوع، أما الموضوع الديمقراطية أنا بتفهم تماما هذا هالوجهة النظر بين ما يمكن نقول إنه الدستور أو
تعديله هو أو إمكانية فتح إعادة انتخاب ديمقراطي بمن المجلس والتصويت إلى ما هالك لرئيس جمهورية بولاية
ثانية هو مصح هيئة في جوهر الديمقراطية هذا بالاقل يعني قناعتي 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 written media (newsgroups, blogs, etc)
- Arabic NLP components for many applications need to account for dialects!
Local Overview: Introduction

- Team
- Problem: Why Parse Arabic Dialects?
- Methodology
- Data Preparation
- Preview of Remainder of Presentation:
  - Lexicon
  - Part-of-Speech Tagging
  - Parsing
Possible Approaches

- Annotate corpora (“Brill Approach”)
- Leverage existing MSA resources
  - Difference MSA/dialect not enormous: can leverage
  - We have linguistic studies of dialects (“scholar-seeded learning”)
  - Too many dialects: even with dialects annotated, still need leveraging for other dialects
  - Code switching: don’t want to annotate corpora with code-switching

OUR APPROACH
Goal of this Work

- Goal of this work: show that leveraging MSA resources for dialects is a viable scientific and engineering option
- Specifically: show that using lexical and structural knowledge of dialects can be used for dialect parsing
- Question of cost ($) is an accounting question
Out of Scope

- Tokenization
- Morphological analyzer (but not a morphological disambiguator)

Speech Effects
  - Repairs and edits
  - Disfluencies
  - Parentheticals
  - Speech sounds

- No standard orthography for dialects
  - Egyptian /mabin\ulhalak$/: mA\ bin\ulhalak$
  - …

- Issue of ASR interface
  - Easy
In Scope

- Deriving bidialectal lexica
- Part-of-speech tagging
- Parsing
Local Overview: Introduction

- Team
- Problem: Why Parse Arabic Dialects?
- Methodology
- Data Preparation
- Preview of Remainder of Presentation:
  - Lexicon
  - Part-of-Speech Tagging
  - Parsing
Arabic Dialects: Computational Resources

- Transcribed speech/transcript corpora
  - Levantine (LDC), Egyptian (LDC), Iraqi, Gulf, ...
- Very little other unannotated text
  - Online: Blogs, newsgroups
  - Paper: Novels, plays, soap opera scripts, ...
- Treebanks
  - Levantine, LDC for this workshop with no funding
  - INTENDED FOR EVALUATION ONLY
- Morphological resources
  - Columbia University Arabic Dialect Project: MAGEAD: Pan-Arab Morphology, only MSA so far (ACL workshop 2005)
  - Buckwalter morphological analyzer for Levantine (LDC, under development, available as black box)
Huge unannoted corpora,
MSA treebank (LDC)
Lexicons,
Morphological analyzers (Buckwalter 2002)
Taggers (Diab et al 2004)
Chunkers (Diab et al 2004)
Parsers (Bikel, Sima’an)
MT system, ASR systems, …
Data Preparation

- 20,000 words of Levantine (Jordanian) syntactically annotated by LDC
- Removed speech effects, leaving 16,000 words (4,000 sentences)
- Divided into development and test data
- Note: NO TRAINING DATA
- Use morphological analysis of LEV corpus as a standin for true morphological analyzer
- Use MSA treebank from LDC (300,000 words) for training and development
- Contributors: Mona Diab, Nizar Habash
Issues in Test Set

- Annotated Levantine corpus used only for development, testing (no training)
- Corpus developed rapidly at LDC (Maamouri, Bies, Buckwalter), for free (thanks!)
- Issues in corpus:
  - 5% words mis-transcribed
  - Some inconsistent annotations
Local Overview: Introduction

- Team
- Problem: Why Parse Arabic Dialects?
- Methodology
- Data Preparation

Preview of Remainder of Presentation:

- Lexicon
- Part-of-Speech Tagging
- Parsing
Bidialectal Lexicons

- Problem:
  - No existing bidialectal lexicons (even on paper)
  - No existing parallel corpora MSA-dialect

- Solution:
  - Use human-written lexicons
  - Use comparable corpora
  - Estimate translation probabilities
Part-of-Speech Tagging

- **Problem:**
  - No POS-annotated corpus for dialect

- **Solution 1: adapt existing MSA resources**
  - Minimal linguistic knowledge
  - MSA-dialect lexicon

- **Solution 2: find new types of models**
Local Overview: Introduction

- Team
- Problem: Why Parse Arabic Dialects?
- Methodology
- Preview of Remainder of Presentation:
  - Lexicon
  - Part-of-Speech Tagging
  - Parsing
Parsing Arabic Dialects: The Problem

- Dialect -

- MSA -

Treebank

Parser

Big UAC

Small UAC

الأولاد كتبوا الأشعار

كتبوا الأشعار
Parsing Solution 1: Dialect Sentence Transduction

- Dialect -
  الاولاد كتبوا الاشعار
  كتابوا الاشعار
  الاولاد
  الاشعار

Translation Lexicon

- MSA -
  كتب الاولاد الاشعار
  كتاب الاشعار
  الاولاد
  الاشعار

Parser

Big LM

Workshop Accomplished
Pre-Existing Resources
Continuing Progress
Parsing Solution 2: MSA Treebank Transduction

- Dialect -

- MSA -

Small LM

Treebank

Parser

Tree Transduction

Workshop Accomplished

Pre-Existing Resources

Continuing Progress
Parsing Solution 3: MSA Grammar Transduction

**- Dialect -**

الاولاد كتبوا الأشعار

كتبوا

الأشعار

**- MSA -**

Probabilistic TAG

Treebank

Probabilistic TAG

Tree Transduction

TAG = Tree Adjoining Grammar

Workshop Accomplished

Pre-Existing Resources

Continuing Progress
What We Have Shown

- Baseline: MSA-trained parser on Levantine
  - Baseline: 53.1%

- This work: a small amount of effort improves
  - Small lexicon, 2 syntactic rules: 60.2%

- Comparison: a large amount of effort for treebanking improves more
  - Annotate 11,000 words: 69.3%
Summary: Introduction

- 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
- Want to quickly develop dialect parsers without need for annotation
- Exploit knowledge of differences MSA/dialects to be able to
Global Overview

- Introduction
- **Student Presentation: Safi Shareef**
- Student Presentation: Vincent Lacey
- Lexicon
- Part-of-Speech Tagging
- Parsing
  - Introduction and Baselines
  - Sentence Transduction
  - Treebank Transduction
  - Grammar Transduction
- Conclusion
Arabic Dialect Text Classification

Student Project Proposal

Advisor: Nizar Habash
Student: Safi Shareef

Columbia University, NY
Johns Hopkins University, MD

August 17, 2005
Background

- **Arabic Diglossia**
  - Standard Arabic: formal, primarily written
  - Arabic Dialects: informal, primarily spoken
  - Differences in phonology, morphology, syntax, lexicon
  - Regional Dialect differences (Iraqi, Egyptian, Levantine, etc.)

- **Spectrum of modern Arabic language forms**
  - Hints toward content
MSA & Dialect mixing within the same text

لا أنا ما بعتق لانه عمليه اللي عم بيعارضوا اليوم تمديد للرئيس لحوده هم اللي طالبوا بالتمديد للرئيس الهرواي وبالتالي موضوع منه موضوع مبني على الأرض لانا بحترم أنه يكون في نظرة ديمقراطية للأمور وأنه يكون في احترام للعبة الديمقراطية وأن يكون في ممارسة ديمقراطية ويعتقد إنه الكل في لبنان أو أكثرية ساحقة في لبنان تريد هذا الموضوع، بس بدي يرجع لحظة على موضوع إنجازات العهد يعني نعم نحكي عن إنجازات العهد لكن هل النظام في لبنان نظام رسمي النظام في لبنان من بعد الطائف ليس نظام رسمي وبالتالي السلطة هي عمليا بين الحكومة مجتمعة والرئيس لحود أثبت خلال ممارسته الأخيرة بأنه لما يكون في شخص مسؤول في منصب معين وانا عشت هذا الموضوع شخصيا بمارستي في موضوع الاتصالات لما بياخذ مواقف صاحبة ضمن خطاب ومبادئ خطاب القسم هو إلى جانبه إما مش مطلوب من رئيس جمهورية هو يكون رئيس السلطة التنفيذية لأنه منه بقي في لبنان ما بعد إتفاق الطائف رئيس السلطة التنفيذية عليه التوجيه عليه إبداء الملاحظات عليه القول ما هو خطأ وما هو صحيح عليه تثمر جهود الوطنية الشاملة كي يظل في مصالحة وطنية كي يظل في توافق ما بين المسلم والمسيحي في لبنان يحتضن أبناء هذا البلد ما يترك المسار يروح باتجاه الخطأ نعم إنه خطاب القسم كان موضوع مبادئ طرحت هو ملتزم فيها اللي مشيوا معه وآمنوا فيها التزموا فيها أنا أثبت خلال الأربع سنوات بالتمارسة الحكومية أنه التزمت فيها ولما التزمنا بهذا الموضوع كان الرئيس لحود إلى جنينا في هذا الموضوع، أما الموضوع الديمقراطي أنا بتفهم تماما هذا هالوجهة النظر بس ما ممكن نقول إنه الدستور أو تعديله هو أو إمكانية فتح إعادة انتخاب ديمقراطي ضمن المجلس والتصويت إلى ما هناك لرئيس جمهورية بولاية ثانية هو مسح هيئة في جوهر الديمقراطية هذا بالأقل يعني قناعتي في هذا الموضوع.
Computational Issues

- Modern Standard Arabic
  - Plethora of resources/applications
  - Textual Corpora
  - Treebanks
  - Morphological Analyzers/Generators

- Arabic Dialects
  - Limited or no resources
  - Many dialects with varying degrees of support
Dialect Detection (Identification)

Motivation

- Create more consistent and robust language models
  - Machine translation
    - e.g. Translate into IRQ in colloquial form

- Application matching
  - What lexicon, analyzer, translation system to use?
  - Dialect ID as additional feature to different applications
    - Information retrieval, information extraction, etc.
## Types of Dialect Classification

- Document-based vs. Word-based
- Single Dialect vs. Multiple Dialect
- Form of Dialect

<table>
<thead>
<tr>
<th>Dimensions of Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Word</strong></td>
</tr>
<tr>
<td>Single Dialect</td>
</tr>
<tr>
<td>Multiple Dialect</td>
</tr>
</tbody>
</table>
Difficulty of Dialect Identification...

- **Research Challenges**
  - Require annotated development and test sets
    - Creating annotating resources (i.e. determining dialect)
  - Other resource requirements:
    - e.g. Word analyzers

<table>
<thead>
<tr>
<th></th>
<th>Single Dialect</th>
<th>Multiple Dialect</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Word</strong></td>
<td>* Hard to annotate</td>
<td>* Harder to annotate</td>
</tr>
<tr>
<td></td>
<td>* Need resources</td>
<td>* Need more resources</td>
</tr>
<tr>
<td><strong>Document</strong></td>
<td>URL annotated Corpora</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Textual resources that originate from known dialectal region</td>
<td></td>
</tr>
</tbody>
</table>
The Problem Being Addressed…

- Document-level Multiple Dialect Classification
  - No Resources exist to identify an Arabic document’s dialect
  - Unannotated Corpora exists!
    - (e.g. news groups, blogs, interviews, etc.)
  - Encompasses single dialect document-level classification
  - Precursor to word-level classification
Proposed Solution

- Develop a text level analyzer to rank Arabic text (at the document level) on likelihood of being LEV, EGP, IRQ, MSA, etc ...

Resources

- Multidialectal corpus annotated by region
  - e.g. use URL of newsgroups
- Dialect-specific wordlists
- Any available word-level applications
  - e.g. morphological analyzer
Arabic Dialect Classification vs. Language Identification

- **Language Identification**
  - Different orthographies
  - Primarily unique vocabulary

- **Arabic Dialect Classification**
  - Not a simple Text Categorization Problem
    - Same orthography
    - Similar word roots
    - Non-uniform text
      - Code-switching
Proposed Approach

Extract Features

Classifier (Trained)

Annotated Corpora

Extracted Features

<table>
<thead>
<tr>
<th></th>
<th>EGY</th>
<th>NEG</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSA</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>N-GRAM</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>--</td>
<td></td>
</tr>
</tbody>
</table>

Annotated Corpora

Extracted Features

MSA

EGY

LEV

...
Global Overview

- Introduction
- Student Presentation: Safi Shareef
- **Student Presentation: Vincent Lacey**
- Lexicon
- Part-of-Speech Tagging
- Parsing
  - Introduction and Baselines
  - Sentence Transduction
  - Treebank Transduction
  - Grammar Transduction
- Conclusion
Statistical Mappings of Multiword Expressions Across Multilingual Corpora

Student Project Proposal

Proposal by

Vincent Lacey
Advisor: Mona Diab
Sponsor: Chin-Hui Lee

Georgia Tech
Columbia
Georgia Tech
First, some motivation:

"Ya be trippin' wit' dat tight truck jewelry."

LEXICON

Ya – Ya; Yes, Okay; You
Be – Be, Are, Is
Trippin’ – Tripping, Falling, High, Crazy
Wit’ – Wits, With
Dat – That
Tight – Constricting; Cool; Nice
Truck – Truck Jewelry
Jewelry – Gold Jewelry

Yes be falling wits that constricting truck jewelry
You be high with that cool truck jewelry
You are crazy with that nice gold jewelry.

-5.439
2.07
-1.34

Yes be
be falling
falling wits
wits that
that constricting
constricting truck
truck jewelry

You be
be high
high with
with that
that cool
tcool gold
gold jewelry

You are
are crazy
are crazy with
with that
that nice
nice gold

gold jewelry

-2.07
0.40
0.65
0.45
0.92
0.69
0.18
0.63

0.89
0.70
0.51
0.92
0.72
0.25
0.63
Lexical Issues

- **Treebank transduction**: MSA->Dialect

- **Sentence transduction & grammar transduction**: Dialect->MSA

- 20% of Levantine words are unrecognized by parsers trained on MSA

- No parallel corpora!
Road Map

- Some Intuition
- Mapping Single Words
- Preliminary Results

- Proposal: Mapping Multiword Expressions
  - Approach
  - Advantages & Applications
  - Work Plan
Some Intuition

Optimists play video games, read magazines and listen to the radio more than do pessimists, while pessimists watch more television…

**Read** the lyrics, **listen**, download and. . .

Who would read or even listen to this stuff??

R(read, listen) = 0.72

R(ler, escuchar) = 0.70

Hoy, con una computadora y un programa especial, una persona ciega puede acceder a la primera biblioteca virtual en lengua hispana para discapacitados visuales, llamada Tiflo libros, y leer--mejor dicho, escuchar--miles de libros por su cuenta.

Lo que me gusta hacer...

**LEER**
ESCUCAR
MUSICA Y
SALIR

R(ler, y) = 0.65
Road Map

- Some Intuition
- **Mapping Single Words**
- **Preliminary Results**

- **Proposal: Mapping Multiword Expressions**
  - Approach
  - Advantages & Applications
  - Work Plan
Mapping Single Words: Spearman

Optimists play video games, **read** magazines and **listen** to the radio more than do pessimists, while pessimists watch more television...

**Read** the lyrics, **listen**, download and... Who would **read** or even **listen** to this stuff??

\[
(R^2) = 1 - \frac{6 \sum d^2}{n^3 - n}
\]

Diab & Finch 2000

Lo que me gusta hacer...
**LEER** **ESCUCHAR** **MUSICA** **Y** **SALIR**
Hoy, con una computadora y un programa especial, una persona ciega puede acceder a la primera biblioteca virtual en lengua hispana para discapacitados visuales, llamada Tiflolibros, y **leer**—mejor dicho, **escuchar**—miles de libros por su cuenta.
Repeat with 3 seed words:

\[
\text{truth} = \begin{pmatrix} 0.4305 \\ 0.5547 \\ 0.7120 \end{pmatrix} \quad \text{verisimilitude} = \begin{pmatrix} 0.4326 \\ 0.5937 \\ 0.6785 \end{pmatrix} \quad \text{golden} = \begin{pmatrix} 0.2279 \\ 0.7218 \\ 0.6534 \end{pmatrix}
\]

\[
\langle \text{truth}, \text{verisimilitude} \rangle = 0.9987
\]

\[
\langle \text{truth}, \text{golden} \rangle = 0.9638
\]

Related work: Knight & Koehn 2002
Mapping Single Words: Cognate Filters

**Before...**

december  |  december  |  december  
family   |  family    |  family    
people   |  people    |  people    
china    |  china     |  china     

\[ lcsr = \frac{\text{longest common substring}}{\text{longest string}} \]

\[ \text{lcsr(} \text{government, gouvernement}) = 10/12 \]

Melamed 1995
Mapping Single Words: Map Reduction

involved → involved
foreign → foreign
policy → policy
resolution → school

Recall: 700%  Precision: 700%
Preliminary Results: Method Comparison

(English-English comparable corpora)

<table>
<thead>
<tr>
<th>Methods</th>
<th>Added Entries</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Similarity</td>
<td>1000</td>
<td>86.4%</td>
</tr>
<tr>
<td>Similarity+LCSR</td>
<td>1000</td>
<td>92.5%</td>
</tr>
<tr>
<td>Similarity+LCSR+MapReduce</td>
<td>841</td>
<td>98.8%</td>
</tr>
</tbody>
</table>
Preliminary Results: Comparable Corpora Analysis

<table>
<thead>
<tr>
<th>English-English Corpora</th>
<th>Precision *</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size (words)</td>
<td>Comparable</td>
</tr>
<tr>
<td>100M</td>
<td>99.7% (889)</td>
</tr>
<tr>
<td>20M</td>
<td>99.2% (825)</td>
</tr>
<tr>
<td>4M</td>
<td>96.3% (719)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Arabic MSA-MSA Corpora</th>
<th>Precision *</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size (words)</td>
<td>Comparable</td>
</tr>
<tr>
<td>100M</td>
<td>99.3% (764)</td>
</tr>
<tr>
<td>20M</td>
<td>98.2% (756)</td>
</tr>
<tr>
<td>4M</td>
<td>94.0% (625)</td>
</tr>
</tbody>
</table>

Comparable: Same genre (“same” newswire), overlapping coverage time
Related: Same genre (different newswire), some overlapping coverage time

*type precision
Road Map

- Some Intuition
- Mapping Single Words
- Preliminary Results

Proposal: Mapping Multiword Expressions
  - Approach
  - Advantages & Applications
  - Work Plan
Approach: Intersecting Sets

First pass:

kicked
the
bucket

Second pass:

die

kicked story die shove off
the of company person die
bucket die pail story conclusion

passed bombings bucket peace kicked
Approach: Synthesis

die → Bombs → Bucket → Kicked the bucket

LM → Kicked
Evaluation

- Using MWE data base at Columbia
- Automated—no human intervention
Advantages & Applications

- No seed lexicon required
- No annotated corpora needed
- *Fast* and extensible

Word Clustering

Cross-lingual information retrieval

Phrase-based machine translation

- many
- issue
- aid
- ireland
Work Plan

- Sources: English/Arabic/Chinese Gigaword
- Aug-Sept: Building initial MWE system
- Sept-Oct: Development testing
- Oct-Dec: Final experiments
Global Overview

- Introduction
- Student Presentation: Safi Shareef
- Student Presentation: Vincent Lacey
- Lexicon (Carol Nichols)
- Part-of-Speech Tagging
- Parsing
  - Introduction and Baselines
  - Sentence Transduction
  - Grammar Transduction
  - Treebank Transduction
- Conclusion
Local Overview: Lexicon

- Building a lexicon for parsing
  - Get the word to word relations
    - Manual construction
    - Vincent Lacey’s presentation (Finch & Diab, 2000)
    - A variant of Rapp (1999)
    - Combination of resources
  - Assign probabilities
- Ways of using lexicons in experiments
Rapp, 1999

<table>
<thead>
<tr>
<th>English corpus</th>
<th>Pig latin corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>People who like to read books are interesting.</td>
<td>e-way ike-lay o-tay ead-ray ooks-bay.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>like</th>
<th>books</th>
</tr>
</thead>
<tbody>
<tr>
<td>are</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>read</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>ike-lay</th>
<th>ooks-bay</th>
</tr>
</thead>
<tbody>
<tr>
<td>ead-ray</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>e-way</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Automatic Extraction from Comparable Corpora

- Novel extensions to Rapp, 1999:
  - Modification: add best pair to dictionary and iterate
  - When to stop? How “bad” is “bad”? 

- English to English corpus: halves of *Emma* by Jane Austen
  - 97% of ~100 words added to dictionary correct
  - 39.5% of other words correct in top candidate
  - 61.5% of other words correct in top 10
Application to LEV-MSA

- Levantine development data & part of MSA treebank:
  - Used words that appeared in both corpora as seed dictionary
  - Held out known words: <10% in top 10
  - Manual examination: sometimes clusters on POS

- Explanation:
  - These are small and unrelated corpora
  - If translation is not in other corpus, no chance of finding it!
  - Levantine: speech about family, MSA: text about politics, news

- Contributors: Carol Nichols, Vincent Lacey, Mona Diab, Rebecca Hwa
Manual Construction

- Simple modification
- Bridge through English
- Manually created

Combination:

Contributors: Nizar Habash
Add Probabilities to Lexicons

- No parallel corpora to compute joint distribution
- Applying EM algorithm using unigram frequency counts from comparable corpora and many-to-many lexicon relations

Contributors: Khalil Sima’an, Carol Nichols, Rebecca Hwa, Mona Diab, Vincent Lacey
Lexicons Used

- Does not rely on corpus specific information
  - Levantine closed class words
  - Top 100 most frequent Levantine words
- Uses info from our dev set:
  - occurrence, POS
  - Combined manual lexicon
  - Combined manual lexicon pruned
    - Leaves only non-MSA-like entries and translations found in ATB
- Transformed lexemes to surface forms using ARAGEN (Habash, 2004)
- Contributors: Nizar Habash, Carol Nichols, Vincent Lacey
## Experiment Variations

<table>
<thead>
<tr>
<th>POS tags</th>
<th>No Lexicon</th>
<th>Small Lexicon</th>
<th>Big Lexicon</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Automatic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gold</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Lexical Issues Summary

Main conclusions:
- Automatic extraction from comparable corpora for Levantine and MSA is difficult
- Using small and big lexicons can improve POS tagging and parsing

Future directions:
- Try other automatic methods (Ex: tf/idf)
- Try to find more comparable corpora
POS Tagging

- Assign parts-of-speech to Levantine words
- Correctly tagged input gives higher parsing accuracies

Assumptions
- Have MSA resources
- Levantine data is tokenized
- Use reduced “Bies” tagset

Contributors: Rebecca Hwa and Roger Levy
Porting from MSA to LEV

- Lexical coverage challenge
  - 80% of word tokens overlap
  - 60% of word types overlap
  - 6% of the overlapped types (10% of tokens) have different tags

- Approaches
  - Exploit readily available resources
  - Augment model to reflect characteristics of the language
Basic Tagging Model: HMM

- Transition distributions: $P(T_i \mid T_{i-1})$
- Emission distributions: $P(W_i \mid T_i)$
- Initial model: MSA Bigram
  - Trained on 587K manually tagged MSA words
Tagging LEV with MSA Model

- **Baselines:** Train on MSA
  - Test on MSA: 93.4%
  - Test on LEV:
    - Dev (11.1K words): 68.8%
    - Test (10.6K words): 64.4%

- **Train on LEV**
  - 10-fold cross validation on LEV Dev: 82.9%
  - Train on LEV Dev, Test on LEV test: 80.2%

- Higher accuracies (~70%) are possible with models such as SVM (Diab et al., 2004)
Naïve Porting

- Assume no change in transitions $P(T_i|T_{i-1})$
- Adapt emission probabilities $P(W|T)$
  - Reclaim mass from MSA-only words
  - Redistribute mass to LEV-only words proportional to unigram frequency
- Unsupervised re-training with EM
- Results on LEV dev:
  - 70.2% without retraining
  - 70.7% after one iteration of EM
  - Further retraining hurts performance
- Result on LEV test: 66.1%
Error Analyses on LEV Dev

- Transition
  - Genre/Domain differences affect transition probabilities
  - Retraining transition probabilities improves accuracy

- Emission
  - Accuracy of MSA-LEV shared words: 84.4%
  - Accuracy of LEV-only words: 16.9%
  - Frequent errors on closed-class words

- Retraining
  - Naïve porting doesn’t give EM enough constraints
Relative proportion of seen/unseen words in Levantine development set

- Count
- Unseen words
- Seen words

Categories: Open-class (NN, VBP, JJ, VBD) and Closed-class (IN, PRP, PRP$)
Tagging accuracy for open-class parts of speech

- Overall accuracy
- Seen-word accuracy
- Unseen-word accuracy

F1

- NN
- VBP
- JJ
- VBD
Exploit Resources

- Minimal linguistic knowledge
  - Closed-class vs. open-class
  - Gather stats on initial and final two letters
    - e.g., Al+ suggests Noun, Adj.
  - Most words have one or two possible Bies tags

- Translation lexicons
  - “Small” vs. “Big”

- Tagged dialect sentences

- Morphological analyzer (Duh&Kirchhoff, 2005)
Tagging Results on LEV Test

<table>
<thead>
<tr>
<th>POS tags</th>
<th>No Lexicon</th>
<th>Small Lexicon</th>
<th>Big Lexicon</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Automatic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gold</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Tagging Results on LEV Test

<table>
<thead>
<tr>
<th></th>
<th>No Lexicon</th>
<th>Small Lexicon</th>
<th>Big Lexicon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive Port</td>
<td>66.6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minimal Linguistic Knowledge</td>
<td>70.5%</td>
<td>77.0%</td>
<td>78.2%</td>
</tr>
<tr>
<td>+100 Tagged LEV Sentences (300 words)</td>
<td>78.3%</td>
<td>79.9%</td>
<td>79.3%</td>
</tr>
</tbody>
</table>

- Baseline: MSA as-is: 64.4%
- Supervised (~11K tagged LEV words): 80.2%
Ongoing Work: Augment Tagging Model

- Distributional methods promising for POS
  - Clark 2000, 2003: completely unsupervised
- We have much more distr. information
  - Some MSA parameters are useful
- LEV words’ *internal* structure constrainable
  - morphological regularities useful for POS clustering (Clark 2003)
Version 1: Simple Morphology

- \( P(W|T) \) determined with character HMM
  - each POS has separate char. HMM
Version 2: Root-Template Morphology

- Character HMM doesn’t capture lots of Arabic morphological structure
- Templates determine open-class POS
POS Tagging Summary

- Lexical coverage is a major challenge
- Linguistic knowledge helps
- Translation lexicons are useful resources
  - Small lexicon offers biggest bang for $$
- Ongoing work: improve model to take advantage of morphological features
Global Overview

- Introduction
- Student Presentation: Safi Shareef
- Student Presentation: Vincent Lacey
- Lexicon
- Part-of-Speech Tagging
- Parsing
  - Introduction and Baselines (Khalil Sima’an)
  - Sentence Transduction
  - Treebank Transduction
  - Grammar Transduction
- Conclusion
Parsing Arabic Dialect

Baselines for Parsing
Parsing Arabic Dialects:
The Problem

- Dialect -

- MSA -

Treebank

Parser

Big UAC

Small UAC

?
Baselines for Parsing LEV

Alternative baseline approaches to parsing Levantine:

- **Unsupervised**: Unsupervised induction
- **MSA-supervised**: Train statistical parser on MSA treebank

Hypothetical:

- **Treebanking**: Train on small LEV treebank (13k words)

Our approach:

- **Without treebanking**: Porting MSA parsers to LEV
  Exploring simple word transduction
Reminder: LEV Data

MSA is Newswire text – LEV is Callhome

For this project, the following strictly speech phenomena were removed from the LEV data (M. Diab):

- EDITED (restarts) and INTJ (interjections)
- PRN (Parentheticals) and UNFINISHED constituents
- All resulting SINGLETON trees

Resulting data:

- Dev-set (1928 sentences) and Test-set (2051 sentences)
- Average sentence length: about 5.5 wds/sen.

Reported results are F1 scores.
Baselines: Unsupervised Parsers for LEV

Unsupervised induction by PCFG [Klein & Manning].

Induce structure for the gold POS tagged LEV dev-set (R. Levy):

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Unsupervised</td>
<td>42.6</td>
<td>–</td>
<td>50.9</td>
<td>–</td>
</tr>
</tbody>
</table>
**Baselines: MSA Parsers for LEV (1)**

MSA Treebank PCFG (R. Levy and K. Sima’an).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>TB PCFG(Free)</td>
<td>63.5</td>
<td>50.5</td>
<td>56.1</td>
<td>34.7</td>
</tr>
<tr>
<td>TB PCFG(+Gold)</td>
<td>71.7</td>
<td>60.4</td>
<td>66.1</td>
<td>49.0</td>
</tr>
<tr>
<td>TB PCFG(+Smooth)</td>
<td>73.0</td>
<td>62.3</td>
<td>66.2</td>
<td>51.6</td>
</tr>
</tbody>
</table>

Most improvement (10%) comes from gold tagging!

**Free**: bare words input
**+Gold**: gold POS tagged input
**+Smooth**: (+Gold) + smoothed model
## Baselines: MSA Parsers for LEV (2)

Gold tagged input:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>TB PCFG (+G+S)</td>
<td>73.0</td>
<td>62.3</td>
<td>66.2</td>
<td>51.6</td>
</tr>
<tr>
<td>Blex.dep. (Bikel)(^1)</td>
<td></td>
<td>60.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treegram (Sima’an)</td>
<td>73.7</td>
<td>62.9</td>
<td>68.7</td>
<td>51.5</td>
</tr>
<tr>
<td>STAG (Chiang)</td>
<td>73.6</td>
<td>63.0</td>
<td>71.0</td>
<td>52.8</td>
</tr>
</tbody>
</table>

**Free POS Tags**

| STAG (Chiang)                | 55.3         |            |              |            |

Treebank PCFG doing as well as lexicalized parsers?

---

\(^1\)Gold POS tags partially enforced (N. Habash).
Train a statistical parser on 13k words LEV treebank. How good a LEV parser will we have?

D. Chiang:
- Ten-fold split LEV-dev-set (90%/10%) train/test sets
- Trained STAG-parser on train, tested on test:
  
  Free tags: $F1 = 67.7$  
  Gold tags: $F1 = 72.6$

Questions:
- Will injecting LEV knowledge into MSA parsers give more?
- What kind of knowledge? How hard is it to come by?
Some Numbers About Lexical Differences

Without morphological normalization on either side.

In the LEV dev-set:

- 21% of word tokens are not in MSA treebank
- 27% of \(\langle word, tag \rangle\) occurrences are not in MSA treebank
The Three Fundamental Approaches

**Sentence:** Translate LEV sentences to MSA sentences

**Treebank:** Translate MSA treebank into LEV

**Grammar:** Translate prob. MSA grammar into LEV grammar

Common to all three approaches: word-to-word translation

Let us try simple word-to-word translation
A Cheap Extension to the Baseline

**Hypothesis:** translating a small number of words will improve parsing accuracy significantly (D. Chiang & N. Habash).

Simple transduction “half-way” to LEV treebank parser
Preview of Baseline Results

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Gold POS Tagged Input</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>STAG (Chiang)</td>
<td>73.6</td>
<td><strong>63.0</strong></td>
<td>71.0</td>
<td>52.8</td>
</tr>
<tr>
<td>Not Tagged Input (Free)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>STAG (Chiang)</td>
<td></td>
<td><strong>55.3</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Global Overview

- Introduction
- Student Presentation: Safi Shareef
- Student Presentation: Vincent Lacey
- Lexicon
- Part-of-Speech Tagging
- Parsing
  - Introduction and Baselines
  - Sentence Transduction (Nizar Habash)
  - Treebank Transduction
  - Grammar Transduction
- Conclusion
Sentence Transduction Approach

- Dialect -

الازلام يحبص الكتاب هدا

Translation Lexicon

- MSA -

لا يحب الرجال هذا العمل

Like

(work) ≠ (men)

Contributors: Nizar Habash, Safi Shareef, Khalil Sima’an
Intuition/Insight

- Translation between closely related languages (MSA/Dialect) is relatively easy compared to translation between unrelated languages (MSA,Dialect/English)

- Dialect-MSA translation is easier than MSA-Dialect translation due to rich MSA resources
  - Surface MSA language models
  - Structural MSA language models
  - MSA grammars
Sentence Transduction Approach

- **Advantages**
  - MSA translation created as a side product

- **Disadvantages**
  - No access to structural information for translation
  - Translation can add more ambiguity for parsing
    - Dialect distinct words can become ambiguous MSA words
      - LEV مين myn ‘who’/
      - MSA من mn ‘from’
Translate dialect sentence to MSA lattice

- Lexical choice under-specified
- Linear permutations using string matching transformative rules
- Language modeling
  - Select best path in lattice

العمل هذا الرجال يحب لا

اللازم بيججو ش شغل هذا هادا

men like not work this
MSA Parsing

- Constituency representation
All along, pass links for dialect word to MSA words.
- Retrace to link dialect words to parse
  - Dependency representation necessary

![Diagram showing MSA Parsing, Language Model, Lattice Translation, and Dialect Sentence.]
Retrace to link dialect words to parse

- Dependency representation necessary
Retrace to link dialect words to parse

- Dependency representation necessary

Dialect

Sentence

ازلام

هادا

بىحبو

ش

شغل

Language

Model

MSA

Parsing

Lattice

Translation

Dialect

Sentence

men

like

not

work

this
### DEV Results

- **Bikel Parser, unforced gold tags, uniform translation probabilities**
  - PARSEVAL P/R/F1

<table>
<thead>
<tr>
<th>Tags</th>
<th>No Lexicon</th>
<th>Small Lexicon</th>
<th>Big Lexicon</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>59.4/51.9/55.4</td>
<td>63.8/58.3/61.0</td>
<td>65.3/61.1/63.1</td>
</tr>
<tr>
<td>Gold</td>
<td>64.0/58.3/61.0</td>
<td>67.5/63.4/65.3</td>
<td>66.8/63.2/65.0</td>
</tr>
</tbody>
</table>

- **POS tagging accuracy**

<table>
<thead>
<tr>
<th>Tags</th>
<th>No Lexicon</th>
<th>Small Lexicon</th>
<th>Big Lexicon</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>71.3</td>
<td>80.4</td>
<td>83.9</td>
</tr>
<tr>
<td>Gold</td>
<td>87.5</td>
<td>91.3</td>
<td>88.6</td>
</tr>
</tbody>
</table>
## TEST vs DEV

- **PARSEVAL P/R/F1**

<table>
<thead>
<tr>
<th></th>
<th>Lexicon</th>
<th>None</th>
<th></th>
<th>Lexicon</th>
<th>Small</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tags</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DEV</td>
<td>TEST</td>
<td></td>
<td></td>
<td>DEV</td>
<td>TEST</td>
</tr>
<tr>
<td>None</td>
<td>55.4</td>
<td>53.5</td>
<td></td>
<td>61.0</td>
<td>57.7</td>
</tr>
<tr>
<td>Gold</td>
<td>61.0</td>
<td>60.2</td>
<td></td>
<td>65.3</td>
<td>64.0</td>
</tr>
</tbody>
</table>

- **POS tagging accuracy**

<table>
<thead>
<tr>
<th></th>
<th>Lexicon</th>
<th>None</th>
<th></th>
<th>Lexicon</th>
<th>Small</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tags</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DEV</td>
<td>TEST</td>
<td></td>
<td></td>
<td>DEV</td>
<td>TEST</td>
</tr>
<tr>
<td>None</td>
<td>71.3</td>
<td>67.4</td>
<td></td>
<td>80.4</td>
<td>74.6</td>
</tr>
<tr>
<td>Gold</td>
<td>87.5</td>
<td>86.6</td>
<td></td>
<td>91.3</td>
<td>89.8</td>
</tr>
</tbody>
</table>
Additional Experiments

- EM translation probabilities
  - Not much or consistently helpful
- Lattice Parsing alternative (Khalil Sima’an)
  - Using a structural LM (but no additional surface LM)
  - No EM probs used
  - PARSEVAL F1 score

<table>
<thead>
<tr>
<th>Tags</th>
<th>Lexicon</th>
<th>None</th>
<th>Lexicon</th>
<th>Small</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DEV</td>
<td>TEST</td>
<td>DEV</td>
<td>TEST</td>
</tr>
<tr>
<td>Gold</td>
<td>62.9</td>
<td>62.0</td>
<td>63.0</td>
<td>61.9</td>
</tr>
</tbody>
</table>
Linear Permutation Experiment

- Negation permutation
  - V $/RP \rightarrow IA/RP V$
- 3% in Dev, 2% in Test
- Dependency accuracy

<table>
<thead>
<tr>
<th>Tags</th>
<th>DEV</th>
<th>TEST</th>
</tr>
</thead>
<tbody>
<tr>
<td>NoPerm</td>
<td>69.6</td>
<td>67.6</td>
</tr>
<tr>
<td>PermNeg</td>
<td>69.7</td>
<td>67.3</td>
</tr>
</tbody>
</table>
Conclusions & Future Plans

- Framework for sentence transduction approach
- 22% reduction on pos tagging error (DEV=32%)
- 9% reduction on F1 labeled constituent error (DEV=13%)

- Explore a larger space of permutations
- Better LMs on MSA
- Integrate surface LM probabilities in lattice parsing approach
- Use Treebank/Grammar transduction parses (without lexical translation)
Global Overview

- Introduction (Owen Rambow)
- Student Presentation: Safi Shareef
- Student Presentation: Vincent Lacey
- Lexicon
- Part-of-Speech Tagging
- Parsing
  - Introduction and Baselines
  - Sentence Transduction
  - Treebank Transduction (Mona Diab)
  - Grammar Transduction
- Conclusion
MSA Treebank Transduction

- Dialect -

Small LM

Treebank

Parser

الإزلام بيحبو ش الشغل هادا

بيبحبو

الإزلام ش الشغل

هادا

- MSA -

Tree Transduction

Treebank
Objective

S
| VP
| | PRT VBP yHb np NP AlrjAl ن NNS AlAzIAm الازلام NP-TPC
| | | NP h*A N AlEml العمل
| | | VP VBP byHbw ن PRT RP $ ن Al$gl N ش نظام
| | | | DT hAdA هدا
| | | | NP

114
Approach

- Structural Manipulations
  - Tree normalizations
  - Syntactic transformations

- Lexical Manipulations
  - Lexical translations
  - Morphological transformations
Resources Required

- MSA Treebank (provided by LDC)

- Knowledge of systematic structural transformations (scholar seeded knowledge)

- Tool to manipulate existing structures (Tregex & Tsurgeon)

- Lexicon of correspondences from MSA to LEV (automatic + hand crafted)

- Evaluation corpus
Tregex (Roger Levy)

S

NP

VP

V didn’t bother him that I showed up

SBAR

that I showed up

regex “it” or “It”

descendant through VP chain

headed by

SBAR = sbar > (VP >+VP (S < (NP = np <<# /^[li]t$/))))

dominates

child-of
Tsurgeon (Roger Levy)

\[ \text{S} \rightarrow \text{NP} \rightarrow \text{VP} \rightarrow \text{S} \]

\[ \text{SBAR} \rightarrow \text{VP} \rightarrow \text{SBAR} \]

\[ \text{V} \rightarrow \text{didn't} \]

\[ \text{VP} \rightarrow \text{bother} \]

\[ \text{NP} \rightarrow \text{him} \]

\[ \text{that I showed up} \]

\[ \text{prune sbar} \]

\[ \text{replace np sbar} \]
Tree Normalizations

Fixing annotation inconsistencies in MSA TB

SBAR → SBARQ
interrogative

Removing superfluous Ss

S S × S S × S S → S S
Syntactic Transformations

- SVO-VSO
- Fragmentation
- Negation
- Demonstrative Pronoun flipping
Syntactic Transformations
Syntactic Transformations
Syntactic Transformations
Lexical Transformations

- Using the dictionaries for finding word correspondences from MSA to LEV {Habash}
  
  - SM: Closed Class dictionary in addition to the 100 most frequent terms and their correspondences
  - LG: SM + open class LEV TB dev set types

- Two types of probabilities associated with entries in dictionary: {Nichols, Sima’an, Hwa}
  
  - EM probabilities
  - Uniform probabilities
Morphological Manipulations

- Replacing all occurrences of MSA VB ‘want’ to NN ‘bd’ and inserting possessive pronoun

- Replacing MSA VB /lys/ by and RP m$”

- Changing VBP verb to VBP b+verb
Experiments

- Tree normalization
- Syntactic transformations
- Lexical transformations
- Morphological transformations
- Interactions between lexical, syntactic and morphological transformations

**Parser**
- Bikel Parser off-shelf

**Evaluation**
- Labeled precision/Labeled recall/F-measure
## Experiment Variations

<table>
<thead>
<tr>
<th>POS tags</th>
<th>No Lexicon</th>
<th>Small Lexicon</th>
<th>Big Lexicon</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>53.2F</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Automatic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gold</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Performance on DevSet

Experimental Conditions

Gold POS

7.7% Error Red.

F-measure

structural

lexical

morph

mix

66
63
60
57
54

Experimental Conditions

60.1
### Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>60.1</td>
<td>60.2</td>
</tr>
<tr>
<td>TNORM+NEG</td>
<td>62</td>
<td>61</td>
</tr>
<tr>
<td>Lex SM+EMprob</td>
<td>61.2</td>
<td>59.7</td>
</tr>
<tr>
<td>MORPH</td>
<td>60.8</td>
<td>60</td>
</tr>
<tr>
<td>Lex SM+EMprob +MORPH</td>
<td>61</td>
<td>59.8</td>
</tr>
<tr>
<td>TNORM+NEG +MORPH</td>
<td>62</td>
<td>60.6</td>
</tr>
<tr>
<td>TNORM+NEG+Lex SM+EM</td>
<td>63.1</td>
<td>61.5</td>
</tr>
<tr>
<td>TNORM+NEG+Lex SM+EM +MORPH</td>
<td>62.6</td>
<td>61.2</td>
</tr>
</tbody>
</table>
Observations

- Not all combinations help
- Morphological transformations seem to hurt when used in conjunction with other transformations
- Difference in domain and genre account for uselessness of the large dictionary
- EM probabilities seem to play the role of LEV language model
- Caveat: Lexical resources even for closed class are created for LEV to MSA not the reverse (25% type deficiency in coverage of closed class items)
Conclusions & Future Directions

- Resource consistency is paramount

Future Directions

- More Error analysis
- Experiment with more transformations
- Add a dialectal language model
- Experiment with more balanced lexical resources
- Test applicability of tools developed here to other Arabic dialects
- Maybe automatically learn possible syntactic transformations?
Global Overview

- Introduction (Owen Rambow)
- Student Presentation: Safi Shareef
- Student Presentation: Vincent Lacey
- Lexicon
- Part-of-Speech Tagging
- Parsing
  - Introduction and Baselines
  - Sentence Transduction
  - Treebank Transduction
  - Grammar Transduction (David Chiang)
- Conclusion
Grammar Transduction

- Dialect -

الأزلام يحبّو ش الشغل هادّا

بيحبو

الأزلام ش الشغل

هادّا

- MSA -

Tree Transduction

TAG = Tree Adjoining Grammar
Grammar Transduction

- Transform MSA parsing model into dialect parsing model
- More precisely: into an MSA-dialect synchronous parsing model
- Parsing model is defined in terms of tree- adjoining grammar derivations

Contributors: David Chiang and Owen Rambow
Tree-Adjoining Grammar

S
  | NP
  |   | VP
  |   |   | V
  |   |   | NP
  |   |   |   | like
  |   |   |   |   | I
  |   |   |   |   | really
  |   |   |   | NP
  |   |   | NP
  |   | NP
  | NP
  | ADVP
  | I
  | really

S
  | NP
  |   | ADVP
  |   | really
  |   | V
  |   | NP
  |   |   | like
  |   |   | NP
  |   | NP
  | NP
  | NP
  | Arabic
  | Arabic
Thus: to transform a TAG, we specify transformations on elementary trees

![Diagram of tree transformations]
Transforming Probabilities

- MSA parsing model is probabilistic, so we need to transform the probabilities too.
- Make transformations probabilistic: this gives $P(T_{\text{Lev}}|T_{\text{MSA}})$
To parse, search for:

\[ \arg \max P(T_{\text{Lev}}) \approx \arg \max P(T_{\text{Lev}}, T_{\text{MSA}}) \]

\[ = \arg \max P(T_{\text{Lev}}|T_{\text{MSA}}) P(T_{\text{MSA}}) \]

given by grammar transformation

learned from MSA treebank
Probability Model

- Full set of mappings is very large, because elementary trees are lexicalized.
- Can backoff to translating unlexicalized part and lexical anchor independently.
Transformations

- VSO to SVO transformation
- Negation:
Transformations

- ‘want’

```
  S  
  |   
  VP  
  |   
  V  S'  
  |   
  >ryd  

  S  
  |   
  VP  
  |   
  N  PRP$  S'  
  |   |   
  bd  y  
```
## Experiments (devtest)

<table>
<thead>
<tr>
<th>POS tags</th>
<th>No Lexicon</th>
<th>Small Lexicon</th>
<th>Big Lexicon</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Automatic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gold</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Recall</td>
<td>Prec</td>
<td>F1</td>
</tr>
<tr>
<td>---------------</td>
<td>--------</td>
<td>-------</td>
<td>------</td>
</tr>
<tr>
<td>Baseline</td>
<td>62.5</td>
<td>63.9</td>
<td>63.2</td>
</tr>
<tr>
<td>Small lexicon</td>
<td>67.0</td>
<td>67.0</td>
<td>67.0</td>
</tr>
<tr>
<td>VSO→SVO</td>
<td>66.7</td>
<td>66.9</td>
<td>66.8</td>
</tr>
<tr>
<td>negation</td>
<td>67.0</td>
<td>67.0</td>
<td>67.0</td>
</tr>
<tr>
<td>‘want’</td>
<td>67.0</td>
<td>67.4</td>
<td>67.2</td>
</tr>
<tr>
<td>negation+‘want’</td>
<td>67.1</td>
<td>67.4</td>
<td>67.3</td>
</tr>
</tbody>
</table>
### Experiments (test)

<table>
<thead>
<tr>
<th>POS tags</th>
<th>No Lexicon</th>
<th>Small Lexicon</th>
<th>Big Lexicon</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Automatic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gold</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Recall</td>
<td>Prec</td>
<td>F1</td>
</tr>
<tr>
<td>----------------</td>
<td>--------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>Baseline</td>
<td>50.9</td>
<td>55.4</td>
<td>53.1</td>
</tr>
<tr>
<td>All, no lexical</td>
<td>51.1</td>
<td>55.5</td>
<td>53.2</td>
</tr>
<tr>
<td>All, small</td>
<td>58.7</td>
<td>61.8</td>
<td>60.2</td>
</tr>
<tr>
<td>All, large</td>
<td>60.0</td>
<td>62.2</td>
<td>61.1</td>
</tr>
</tbody>
</table>
Further Results

- Combining with unsupervised POS tagger hurts (about 2 points)
- Using EM to reestimate either $P(T_{Lev}|T_{MSA})$ or $P(T_{MSA})$
  - no lexicon: helps first iteration (about 1 point), then hurts
  - small lexicon: doesn’t help
Conclusions

- Syntactic transformations help, but not as much as lexical.

Future work:
- transformations involving multiple words and syntactic context
- test other parameterizations, backoff schemes
Global Overview

- Introduction
- Student Presentation: Safi Shareef
- Student Presentation: Vincent Lacey
- Lexicon
- Part-of-Speech Tagging
- Parsing
  - Introduction and Baselines
  - Sentence Transduction
  - Treebank Transduction
  - Grammar Transduction
- Conclusion (Owen Rambow)
Accomplishments

- Created software for acquiring lexicons from comparable corpora
- Investigated use of different lexicons in Arabic dialect NLP tasks
- Investigated POS tagging for dialects
- Developed three approaches to parsing for dialects, with software and methodologies
Summary: Quantitative Results

- POS tagging
  - No lexicon to small lexicon: 70% to 77%
  - Small lexicon to small lexicon with in-domain information: 77% to 80%

- Parsing
  - No lexicon to small lexicon: 63.2% to 67%
  - Small lexicon to small lexicon with syntax: 67% to 67.3%
  - Train on 10,000 trebanked words: 69.3%
Resources Created

- Lexicons:
  - Hand-created closed-class, open-class lexicons for Levantine

- POS Tagging:
  - Software for adapting MSA tagger to dialect

- Parsing:
  - Sentence-transduction & parsing software
  - Tree-transformation software
  - Synchronous grammar framework

- Treebanks
  - Transduced dialect treebank
Future Work

- Improve reported work
  - Comparable corpora for Arabic dialects
  - Improve POS results
  - Explore more tree transformations for grammar transduction, treebank transduction
  - Include structural information for key words
- Combine leveraging MSA with use of small Levantine treebank
  - Already used in POS tagging
  - Combine transduced treebank with annotated treebank
  - Augment extracted grammar with transformed grammar