Parsing Arabic Dialects

Progress Report

(- Week 2-)
Overview

• Baselines
• Infrastructure
• Sentence Transduction
• Treebank Transduction
• Grammar Transduction
• Issues
Parsing Arabic Dialects: The Problem

- Dialect -

الأولاد كتيو الاشعار

؟

كتبو

الأولاد الاشعار

- MSA -

Treebank

Parser

Big UAC
## Baselines: Unsupervised, MSA Parsers on LEV

**NOTE:** not everything is always comparable

<table>
<thead>
<tr>
<th>Method</th>
<th>Sents</th>
<th>Recall</th>
<th>Precision</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unsupervised</td>
<td>&lt;10</td>
<td>62</td>
<td>38</td>
<td>47</td>
</tr>
<tr>
<td>Chiang</td>
<td>&lt;200</td>
<td>36</td>
<td>42</td>
<td>39</td>
</tr>
<tr>
<td>Treegram (Sima’an)</td>
<td>all</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bikel</td>
<td>&lt;100</td>
<td>41</td>
<td>45</td>
<td>43</td>
</tr>
</tbody>
</table>
Infrastructure: Corpora

• Created dev/train/test sets for the MSA, LATB data sets
• Prepared the data in MSA Gigaword and the other Levantine data sets for language modeling
• Tree graphing
• Mapping to linguistically motivated dependency structure
Infrastructure: Lexical, Computational Resources

- LEV-MSA dictionary building continues
- Prepared noisy dictionary that uses English as bridge language
- Simulated Levantine morphology through extraction of analyses from LATB
- Re-vamped Diab’s Arabic SVMtools to deal with the MSA and Levantine data
- Implemented tree transformation engine and treebank search engine (better than tgrep...)
Infrastructure: Morphology

- Developing a structure for a multidialectal lexicon of Arabic along the lines of the multilingual lexicon proposal of Cahill & Gazdar 1995
- Lexical entries are a triple `<Root,Semantics,MorphClass>`, where each of these portions of the triple can be inherited from a specification that ranges from covering a single dialect to covering all dialects.
Infrastructure: Unannotated Corpora

- **Goal**: Expand Lexicon to improve word to word translation coverage
- **Method 1**: Similar to Rapp, 1999
  - Find words that have similar co-occurrences with known words in seed dictionary (size of seed dictionary will affect performance)
- **Method 2**: Similar to Diab & Finch, 2000
  - Pick a subset of words from each language to compare co-occurrence vectors with all words in the subset
Proposed Solution 1: Dialect Sentence Transduction

- Dialect -
  الاولاد كتبوا الآشعاء
  - Translation Lexicon -
  كتب
  الاولاد الاشاعر

- MSA -
  كتب الاولاد الاشاعر
  كتب
  الاولاد الاشاعر
  - Parser -
  Big LM

- Workshop Accomplished
- Existing Resources
- Continuing Progress
Sentence Transduction

- Completed the implementation end-to-end for single sentences
  - Overgenerative translation
  - No permutations currently
  - LM pruning
  - Integration with parser
  - “Bread crumbing”
  - Simple projection of MSA parse unto LEV sentence
Proposed Solution 2: MSA Treebank Transduction
Proposed Solution 3: MSA Grammar Transduction

- Dialect -
  
  Probabilistic TAG
  
  Parser
  
- MSA -
  
  Treebank
  
  Probabilistic TAG
  
  Tree Transduction
  
  TAG = Tree Adjoining Grammar

Workshop Accomplished
Existing Resources
Continuing Progress
Grammar Transduction

- Use hand-written rules and tree transformation engine to transform MSA grammar into dialect grammar
- Issue: no dialect treebank
- Idea: use probabilities estimated from MSA treebank
Grammar Transduction: Parsing by Translation

- MSA and dialect grammars form synchronous grammar

DIAGRAM

- Idea: use synchronous grammar to translate dialect into MSA
- Find the best translation into MSA and take the dialect tree produced along the way
Grammar Transduction Picture

P(β|η)

P(β'|β)

MSA grammar

Dialect grammar
Grammar Transduction: Synchronous Grammars

- **Stochastic synchronous grammar**
  - Rules in grammar consist of \((r, r', \diamond)\)
    - \(r\): rule in MSA grammar
    - \(r'\): rule in dialect grammar
    - \(\diamond\): bijective relation between nonterminals in RHSs of \(r, r'\)
  - \(P(X \rightarrow س, X' \rightarrow س' | X\diamond X')\)
    where \(X \rightarrow س\) is MSA and \(X' \rightarrow س'\) dialect

- **Approximate**:
  - \(P(X \rightarrow س, X' \rightarrow س' | X\diamond X') \approx P(X' \rightarrow س' | X \rightarrow س) \times P(X \rightarrow س | X)\)
Grammar Transduction: Estimating Parameters

- Reminiscent of Hidden Markov Model:
  - Transitions are the MSA grammar derivations
  - Emissions are the translation

- Extract grammar and estimate transition probabilities from MSA data

- Create synchronous grammar using handwritten tree transduction rules, estimate initial “emission” probabilities by guessing

- Re-estimate either or both by EM or Gibbs sampling on unannotated dialect data
Grammar/Treebank Transduction: Where We Are

- Synchronous TAG implementation: done
- EM of emission probabilities for synchronous TAG: done
- EM training for dialect PCFGs is ready (using Lopar): tested on a toy example
- LM rescoring on MSA or dialect side: not started yet
Gibbs Sampling

- Basic idea: get a joint distribution by repeatedly sampling conditional distributions
- Joint distribution: over values of parameters and data, both observed and unobserved; observed data fixed
- Much like EM, but advantage: distribution over possible values of parameters instead of just a single parameter value; can spot multiple modes
Gibbs Sampling: Dialect Parsing

- Application to parsing dialect:
  - observed data: dialect sentences, MSA trees
  - hidden data: dialect trees
  - parameters: emission probabilities (and perhaps language model parameters as well)

- Best fit with treebank transfer method: get dialect treebank as a part of the sampling process
Gibbs Sampling: Possible Issues

- Disadvantages:
  - May be computationally too expensive
  - Large parameter space
  - May have no clear global optimum (in which case EM would be fine)

- Potential exploration: only sample for a subset of parameters
Plans for Coming Week

- Baselining for all three approaches
- Sentence transduction:
  - More complex permutations
  - Efficient implementation
- Treebank, grammar transduction:
  - More tree transformation rules
  - Implement and test EM
  - Small trial for Gibbs sampling
Issues

• How do we deal with speech aspects? Parentheticals, edits, etc – marked in treebank
• Evaluation issues