بسم الله الرحمن الرحيم

Parsing Arabic Dialects

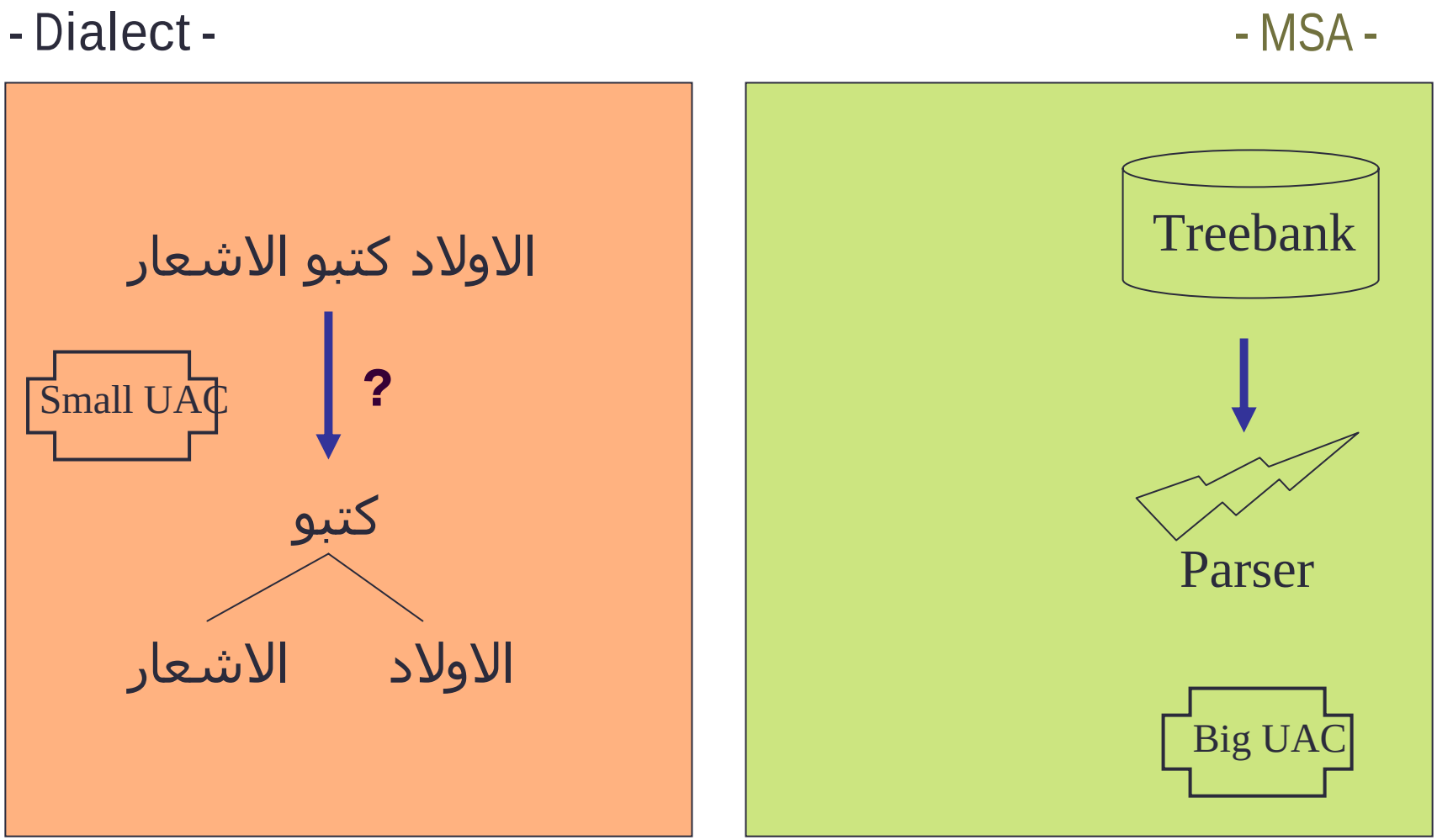
Progress Report

(- Week 2-)

Overview

- Baselines
- Infrastructure
- Sentence Transduction
- Treebank Transduction
- Grammar Transduction
- Issues

Parsing Arabic Dialects:
The Problem



Baselines: Unsupervised, MSA Parsers on LEV

NOTE: not everything is always comparable

Method	Sents	Recall	Precision	F
Unsupervised	<10	62	38	47
Chiang	<200	36	42	39
Treegram (Sima'an)	all			
Bikel	<100	41	45	43

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…)

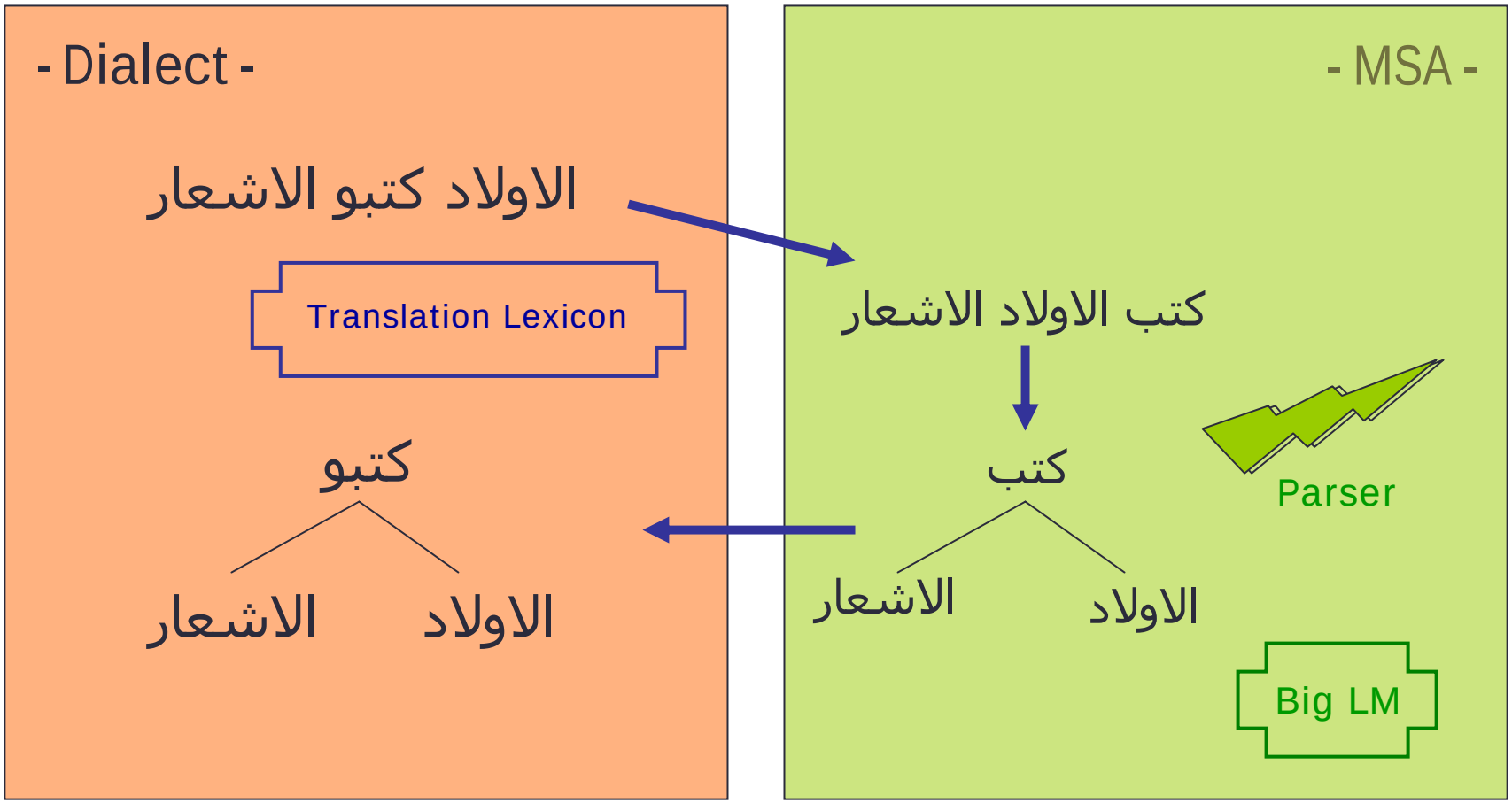
Infrastructure: Morphology

- 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-occurence vectors with all words in the subset

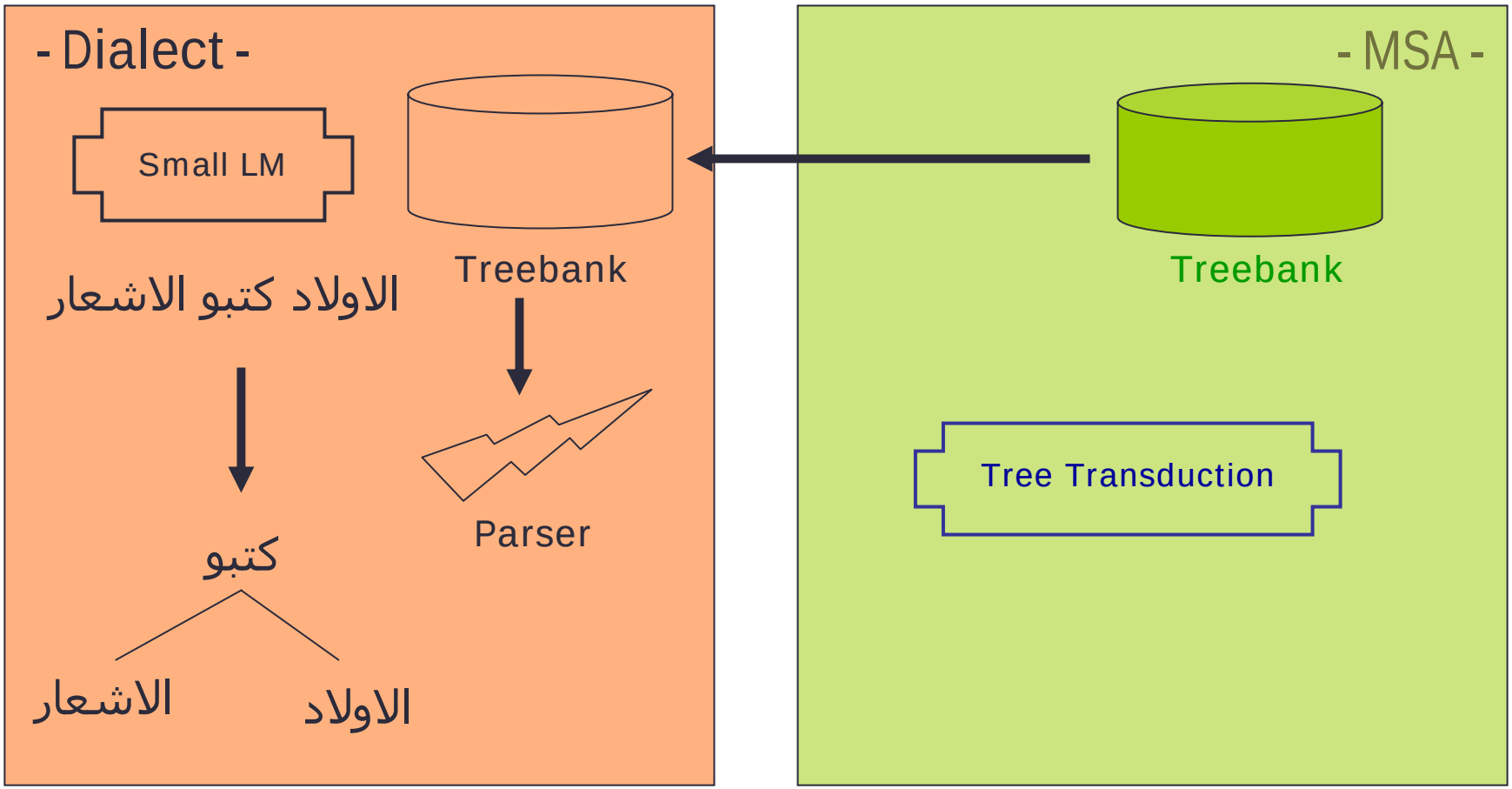
Proposed Solution 1:
Dialect Sentence Transduction



Sentence Transduction

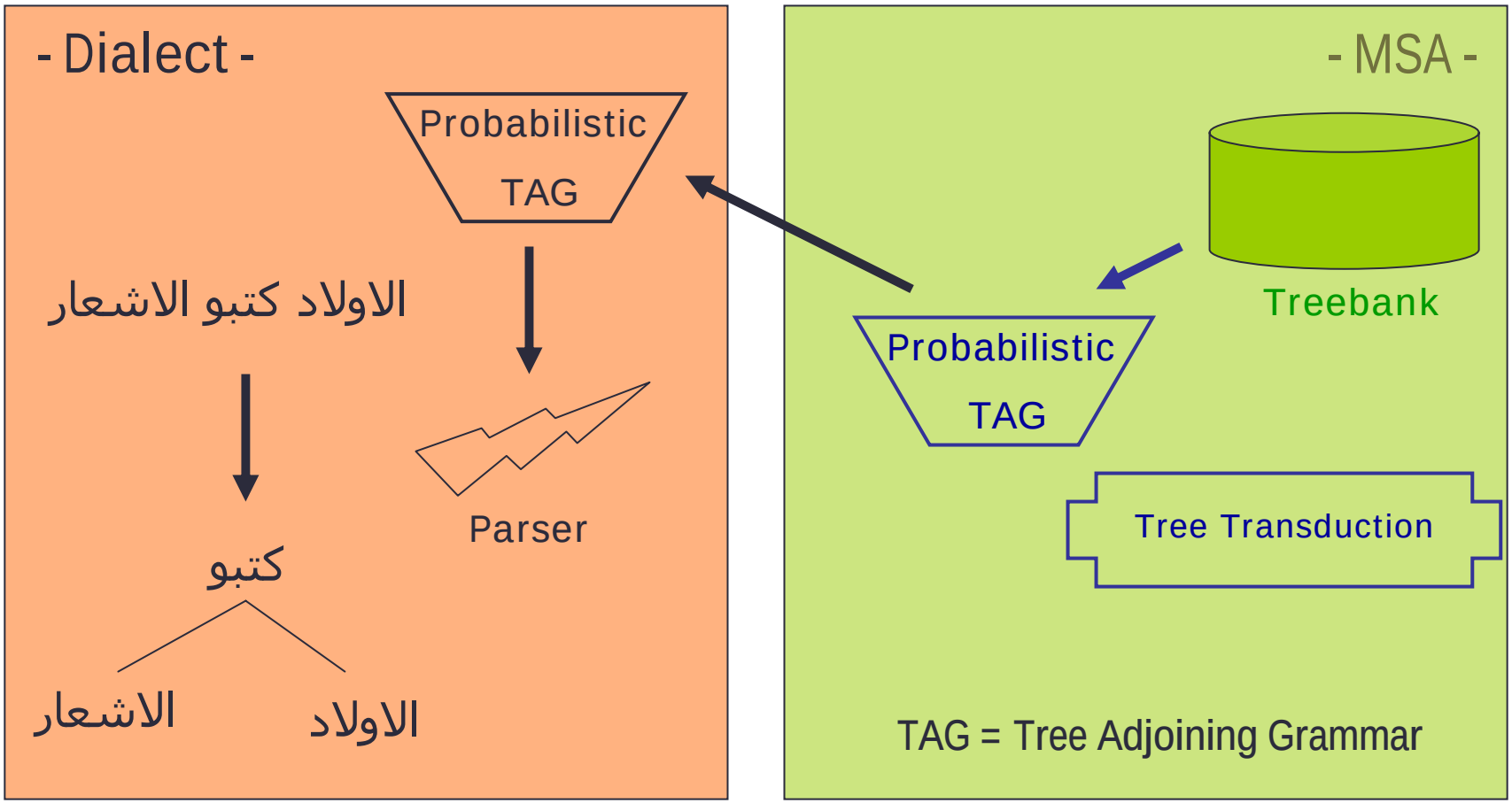
- Completed the implemetation end-to-end for single sentences
  - Overgenerative translation
  - No permutions currently
  - LM pruning
  - Integration with parser
  - “Bread crumbing”
  - Simple projection of MSA parse unto LEV sentence

Proposed Solution 2:
MSA Treebank Transduction



Workshop Accomplished | Existing Resources | Continuing Progress

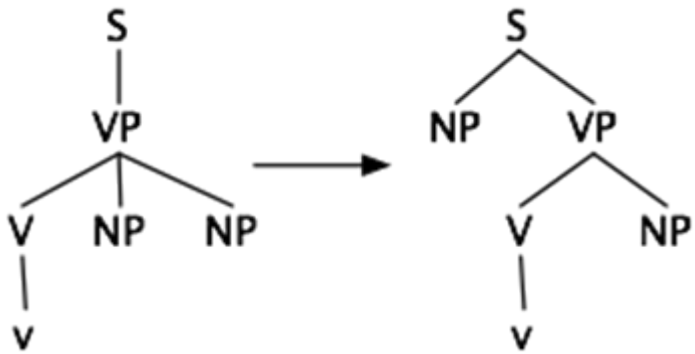
Proposed Solution 3:
MSA Grammar Transduction



Workshop Accomplished | Existing Resources | Continuing Progress

Grammar Transduction

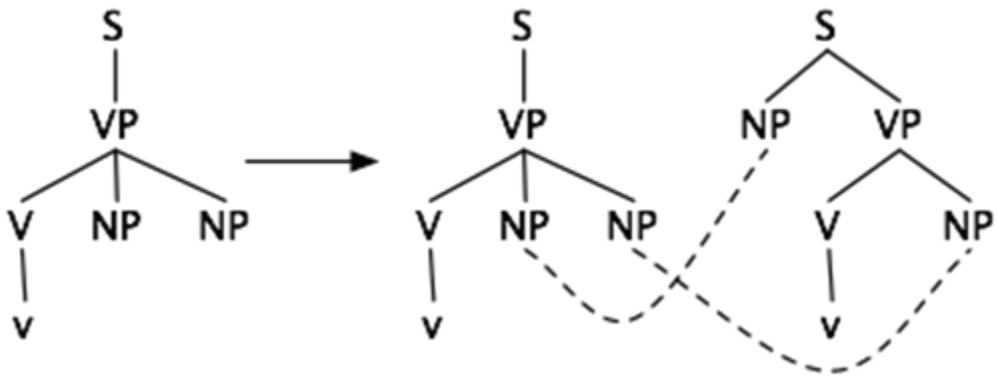
- Use hand-written rules and tree transformation engine to transform MSA grammar into dialect grammar



- Issue: no dialect treebank
- Idea: use probabilities estimated from MSA treebank

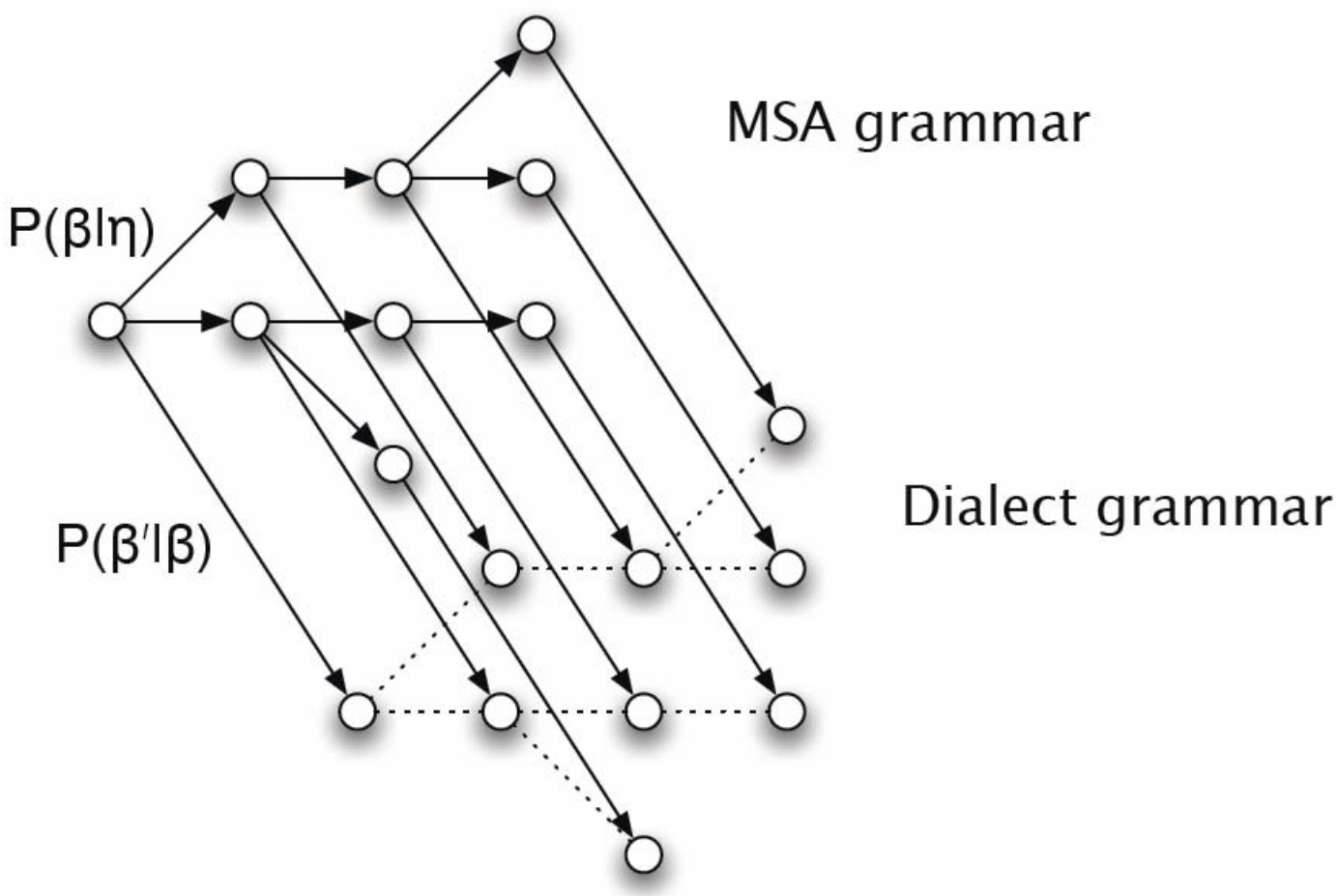
Grammar Transduction: Parsing by Translation

- MSA and dialect grammars form synchronous grammar



- 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



Grammar Transduction: Synchronous Grammars

- Stochastic synchronous grammar
  - Rules in grammar consist of (r, r′, ◊)
    - r: rule in MSA grammar
    - r′: rule in dialect grammar
    - ◊: bijective relation between nonterminals in RHSs of r, r′
  - $P(X \rightarrow$ س, $X' \rightarrow$ س' $\mid X \Diamond X')$
    where $X \rightarrow$ س is MSA and $X' \rightarrow$ س' dialect
- Approximate:
  - $P(X \rightarrow$ س, $X' \rightarrow$ س' $\mid X \Diamond X')$
    $\approx P(X' \rightarrow$ س' $\mid X \rightarrow$ س$) \times P(X \rightarrow$ س $\mid X)$

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