Motivation

Multilingual cues for grammar induction:

**English:** 
I saw the student from MIT

**Urdu:**
I MIT of student saw

Idea:
Learn phrase structure from systematic variation across languages.

Reading: Snyder et al. (2009).
Multilingual cues for grammar induction:

**English:**  
I saw *the student [from MIT]*

**Urdu:**  
I [[MIT of ] student] saw

**Idea:** learn phrase structure from systematic variation across languages.

---

**Tree Alignment**

For trees $T_1$ and $T_2$, an alignment tree $A$ is obtained as follows:

1. insert empty nodes into $T_1$ and $T_2$ and swap sibling order, until they are isomorphic;
2. overlay the resulting trees $T'_1$ and $T'_2$ to obtain $A$.

---

**Tree Alignment**

Two trees (i) with two possible alignments (ii) and (iii):

(ii) involves swapping of siblings and aligns all nodes, (iii) involves insertion of empty nodes ($\lambda$) and aligns only two (arrows).
Introduction

Parsing Model

Evaluation

Motivation

Tree Alignment

Linguistic example:

[Diagram of tree alignment between English and Urdu]

Frank Keller Natural Language Understanding 7

Generative Process

Observed:

Hypothesize alignment trees that best explain:
- frequent POS sequence pairs;
- lexical alignments.

Model closely based on CCM (Klein and Manning 2002).

Parameters (multinomials), with $i \in \{1, 2\}$:
- $\pi^C_i$: over constituent yields of language $i$;
- $\pi^D_i$: over constituent contexts of language $i$;
- $\phi^C_i$: over constituent yields of language $i$;
- $\phi^D_i$: over constituent contexts of language $i$;
- $\omega$: over pairs of constituent yields from the two languages.

Frank Keller Natural Language Understanding 8

Generative Process

Draw $(T_1, T_2, A)$ from a uniform distribution (where $T_1, T_2$ are binary trees and $A$ is an alignment tree).

For each unaligned node in $A$ draw a constituent yield according to $\pi^C_i$ and a constituent context according to $\phi^C_i$.

Frank Keller Natural Language Understanding 9
For each aligned node in $A$ draw a pair of constituent yields according to $\pi_1^C \cdot \pi_2^C \cdot \omega$ and constituent contexts according to $\phi_i^C$.

For each span in $T_i$ not dominated by a node, draw a distituent yield according to $\pi_i^D$ and a distituent context according to $\phi_i^D$.

Draw word-level alignments consistent with Giza scores, according to a uniform distribution.

We want to predict tree pairs $(T_1, T_2)$ which maximize $P(T_1, T_2|s_1, s_2, g)$, where $s_1$ and $s_2$ are the POS sequences in the two languages and $g$ is the Giza word alignment.

For this we would need to marginalize over all $\pi$, $\phi$, and $\omega$ values, and over the alignment trees $A$.

This is intractable; use Gibbs sampling instead (will be discussed in future lectures).
Hard to sample aligned tree pair \((T_1, T_2, A)\) (need to sum over all subtrees and all their alignments). Instead:

- use proposal distribution \(Q\), which assumes no nodes are aligned, to separately sample \(T_1^*, T_2^*\);
- accept with probability (Metropolis-Hastings):
  \[
  \min\{1, \frac{P(T_1^*, T_2^*)Q(T_1, T_2)}{P(T_1, T_2)Q(T_1^*, T_2^*)}\}
  \]
- conditionally sample tree alignment \(A|T_1, T_2\).

Use dynamic programming to efficiently compute \(P(T_1, T_2)\) and \(A|T_1, T_2\).

Hypothesis: bilingual data improves monolingual grammar induction:

- input: bilingual POS sequences with Giza alignments;
- output: binary tree bracketings;
- evaluate bracket precision, recall, F-measure, on held-out monolingual test data.

Baseline: (Bayesian) CCM (Klein and Manning 2002)

Bilingual corpora used for training and testing:

- Korean-English Treebank: 5,000 sentences;
- Urdu translation of WSJ: 4,300 sentences: no Urdu gold brackets;
- English-Chinese Treebank: 3,850 sentences.

Evaluate on various maximum sentence lengths (5–30).
Average improvement across all scenarios:
- Precision: +10
- Recall: +8
- F-measure: +9

Average reduction in error relative to binary tree oracle: 19%.

Summary

Key idea: use bilingual cues to learn better unsupervised monolingual models of grammar;
- adapt Tree Alignment to probabilistic setting:
  - discover partial shared structure;
  - allow language-specific divergence;
  - computationally tractable;
- achieve improved performance on five corpora, across all sentence lengths.