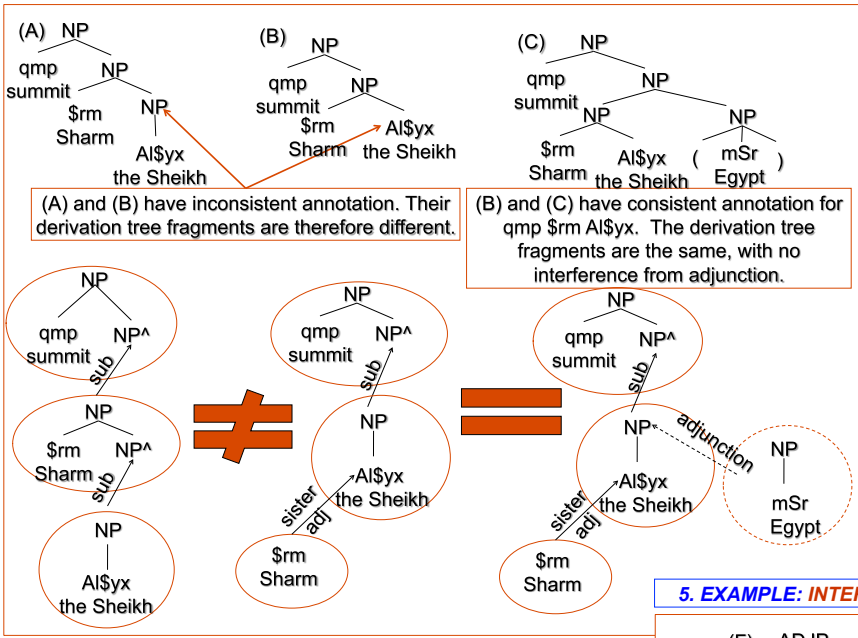
Further Developments in Treebank Error Detection Using Derivation Trees

Seth Kulick, Ann Bies, Justin Mott
{skulick,bies,jmott}@ldc.upenn.edu
Linguistic Data Consortium, University of Pennsylvania

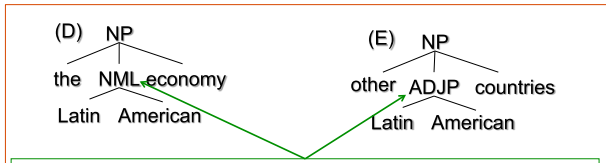
1. GOAL: AUTOMATICALLY DISCOVER INCONSISTENCIES IN TREEBANK ANNOTATION

- Improved detection using "derivation tree fragments" to compare annotations of "nuclei" – Strings of words with possibly inconsistent annotations. Builds upon DECCA (Dickinson & Meurers, 2003).
  - Tree Adjoining Grammar-based decomposition – Each sentence has a derivation tree composed of elementary trees.
  - Derivation Tree Fragment – Restriction of the derivation tree to only those elementary trees with words in the nucleus.
- "Internal" relation check and "external" relation check for different kinds of inconsistencies.

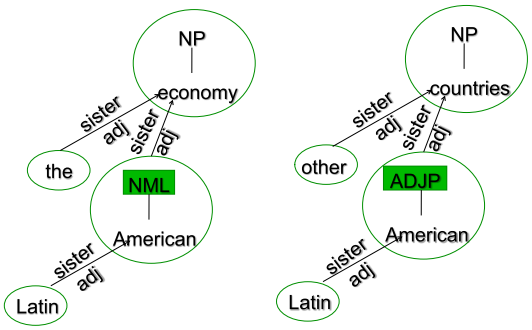
2. EXAMPLE: INTERNAL CHECK IN ARABIC TREEBANK (ATB)



3. EXAMPLE: EXTERNAL CHECK IN ONTONOTES

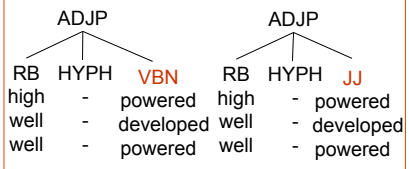


The internal structure of "Latin American" is the same but they have different node labels in the same syntactic context.

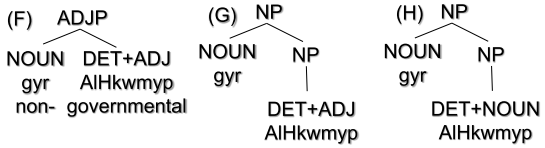


4. IDENTIFYING PATTERNS OF INCONSISTENCY IN INTERNAL CHECK

- Each nucleus instance has one of a finite number of derivation tree fragments.
- Each annotation inconsistency can be characterized by the set of derivation tree fragments for the instances.
- Allows us to sort the results by patterns of inconsistency, e.g.:
  - 9 other nuclei pattern the same as "$rm Al$yx" nucleus.
  - "Well-developed" and "well-powered" pattern the same as "high-powered" in OntoNotes 4.0.



5. EXAMPLE: INTERNAL AND EXTERNAL CHECK TOGETHER



- Difference between (F) and (G) & (H) found by external check (ADJP vs. NP).
- Difference between (G) and (H) found by internal check (DET+ADJ vs. DET+NOUN).

6. EVALUATION

Annotation inconsistencies reported for the ATB

Check	Nuclei found	Non-duplicate nuclei found	Types of inconsistency
Internal	9984	4272	1911
External	191	unknown	n/a

Annotation inconsistencies reported for OntoNotes

Check	Nuclei found	Non-duplicate nuclei found	Types of inconsistency
Internal	3609	3012	1186
External	859	unknown	n/a