

Portable Divergence Unraveling

The Case of Hindi

Nitin Madnani, Necip Fazil Ayan, Bonnie Dorr, Nizar Habash & Christof Monz

What is DUSTER?

- Structural differences - divergences - occur frequently between language pairs (1 out of every 3 sentences).
- Dealing with these divergences is a challenging task for existing statistical word-alignment tools.
- DUSTER (Divergence Unraveling for Statistical Translation) - a rule based framework - systematically identifies and "unravels" common divergence types between a given language pair.
- Achieved by transforming the English sentence into a pseudo-English form (E') which matches the physical form of the foreign language more closely.
- This increases the likelihood of one-to-one correspondences between the words and, hence, achieves improved word alignment between the two languages.
- Improved alignments can be used for projection of dependency trees in another language to serve as input for training parsers for that language - especially useful if the language is resource-poor.

Figure 1. Examples of Divergences between English and Hindi

Divergence Type	English	E'	Hindi
Light Verb	make cuttings	wound	काटा
Manner	The land mourns	The land stays mourning	धरती रोती रहती है
Structural	envy him	envy PREP him	उस से जलता है
Categorical	I am afraid	to-me fear be	मुझे डर है
Head-Swapping	is valued at 4 rupees	value be 4 rupees	मूल्य चार रुपये है
Thematic	I am pained	to-me pain they	मुझे दुःख देते हैं

Figure 2 . Examples of DUSTER Universal Transformation Rules

Light Verb :

A. Expansion: $[V_i(\text{PsychV}) \text{ Arg}_j] \rightarrow [V(\text{LightVB}) \text{ Arg}_j \text{ N}_i]$

Example: "I fear" \rightarrow "I have fear"

B. Contraction: $[V(\text{LightVB}) \text{ Arg}_i \text{ Adj}_j] \rightarrow [V_j(\text{DirectionV}) \text{ Arg}_i]$

Example: "our hand is high" \rightarrow "our hand heightened"

Manner:

C. Expansion: $[V_i \text{ Arg}_j] \rightarrow [V(\text{MotionV}) \text{ Arg}_j \text{ Modifier}_i]$

Example: "I teach" \rightarrow "I walk teaching"

D. Contraction: $[V(\text{ChangeOfStateV}) \text{ Arg}_i \text{ Modifier}_j] \rightarrow [V_j(\text{DirectionV}) \text{ Arg}_i]$

Example: "he turns again" \rightarrow "He returns"

Example: "he turns again" \rightarrow "He returns"

Structural :

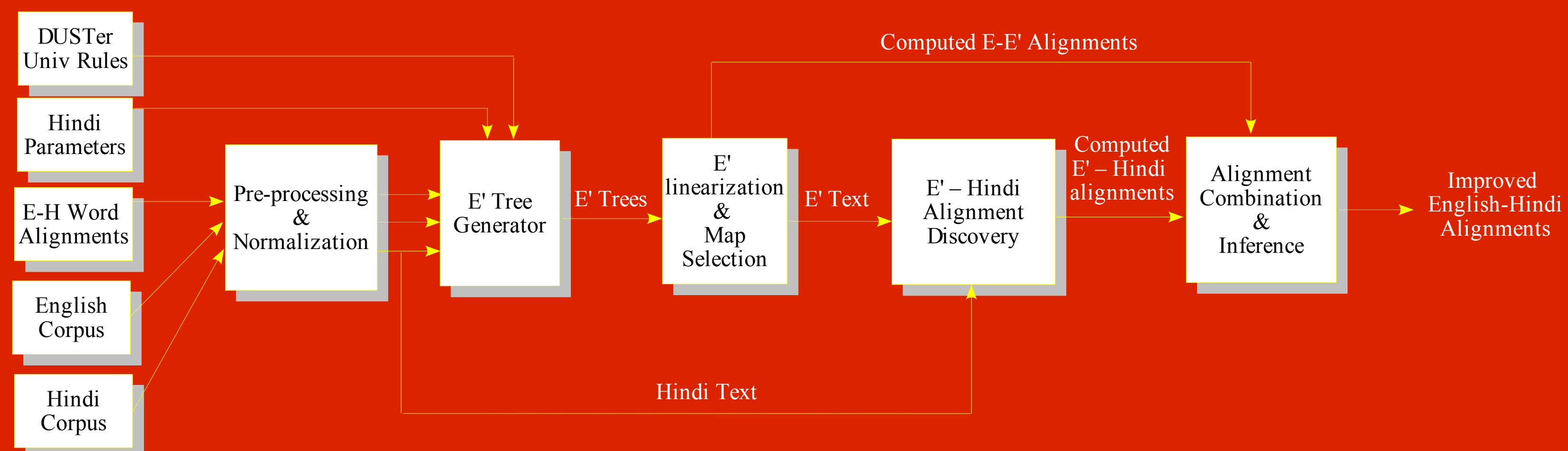
E. Expansion: $[V_i \text{ Arg}_j \text{ Arg}_k] \rightarrow [V_i \text{ Arg}_j \text{ P}(\text{Oblique}) \text{ Arg}_k]$

Example: "I forsake thee" \rightarrow "I forsake of thee"

F. Contraction: $[V_i \text{ Arg}_j \text{ P}(\text{Oblique}) \text{ Arg}_k] \rightarrow [V_i \text{ Arg}_k \text{ Arg}_j]$

Example: "I search for him" \rightarrow "I search him"

Figure 3. DUSTER System Architecture



Adapting DUSTER to Hindi

- DUSTER ported to Hindi during the DARPA TIDES-2003 Surprise Language Exercise 2003.
- Divergences categories based on large-scale multilingual analysis (Spanish, Arabic, and then Hindi)
- Divergence types filled with Hindi examples using surprise-language data (BBC, EMILE and the electronic Bible) [Figure 1]
- Hindi divergences accommodated by 21 existing rules. No new rules added [Figure 2].
- For each rule, LHS represents English and RHS represents Hindi.
- Each rule lexically parameterized with a set of pre-specified lexical items, serving as lexical triggers for rule applicability. For example, V(LightVB) in rule B can only instantiate as one of *do*, *be*, *take*, *give*, *have* and *put*.
- Parameters set rapidly – involved human translation of 16 English parameter settings to their Hindi counterparts.
- Entire porting process well under 3 person days [Figure 4].
- Time trade off exhibited due to morphological richness (Arabic – High, Hindi – Low, Chinese – Negligible).

Figure 4. Times for human porting of DUSTER to Hindi
[Arabic and Chinese times provided for comparison]

Task	Hindi	Arabic	Chinese
Parameter Setting	3.7 hours = 0.5 days	3.3 hours = 0.4 days	17.15 hours = 2.1 days
Morph Specification	8 hours = 1 day	16 hours = 2 days	0 hours = 0 days
Total time	1.5 days	2.4 days	2.1 days