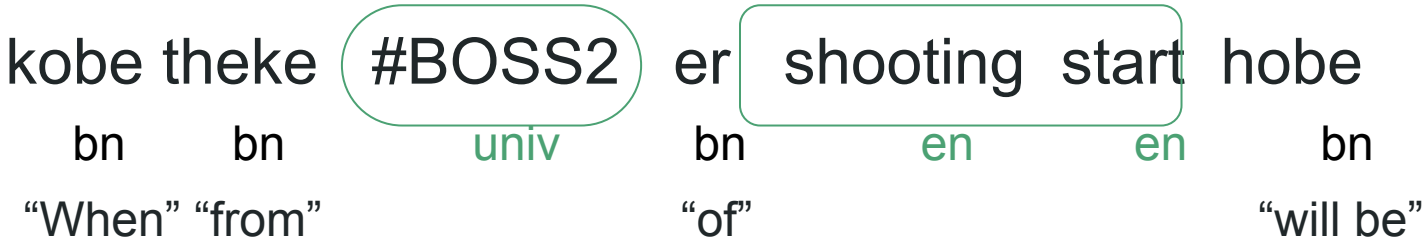
Dependency Parser for Bengali-English Code-Mixed Data enhanced with a Synthetic Treebank

Urmi Ghosh, Dipti Misra Sharma and Simran Khanuja

LTRC, IIIT-H, India

Code-Mixing

- mixing of various linguistic units
- from two (or more) languages
- within a sentence



Bengali-English CM

Bengali

- the second most widely spoken language in India after Hindi (Bhatia, 1982)
- the official and national language of Bangladesh
- 261 million speakers (Ethnologue, 2018)

- Language Identification (Das and Gambäck, 2014)
- POS tagging (Jamatia et al., 2015)
- Dependency parser (Bhat, 2018) - Hindi-English!

Similarities with Hi-EN

Hindi + English

SOV SVO

Bengali + English

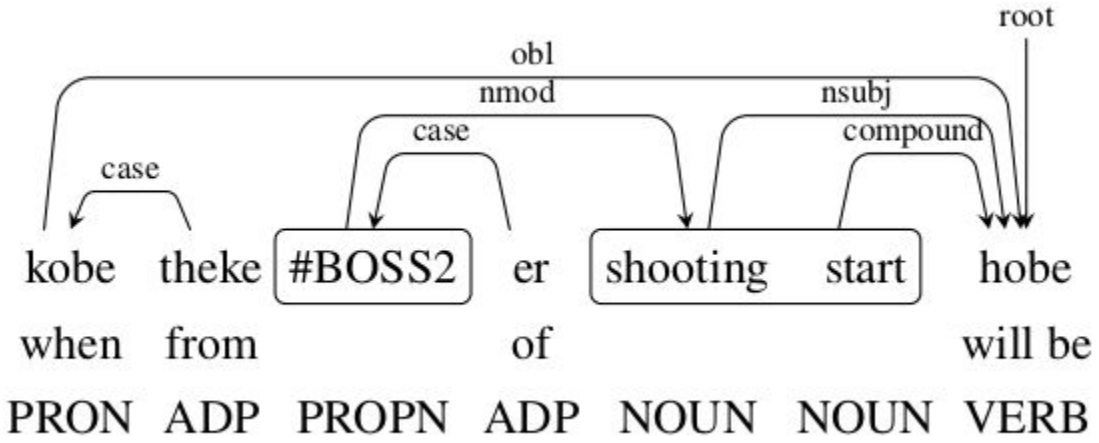
- dirty hands ke use se bache
- dirty hands era use ediye chalun

Data Preparation and Annotation

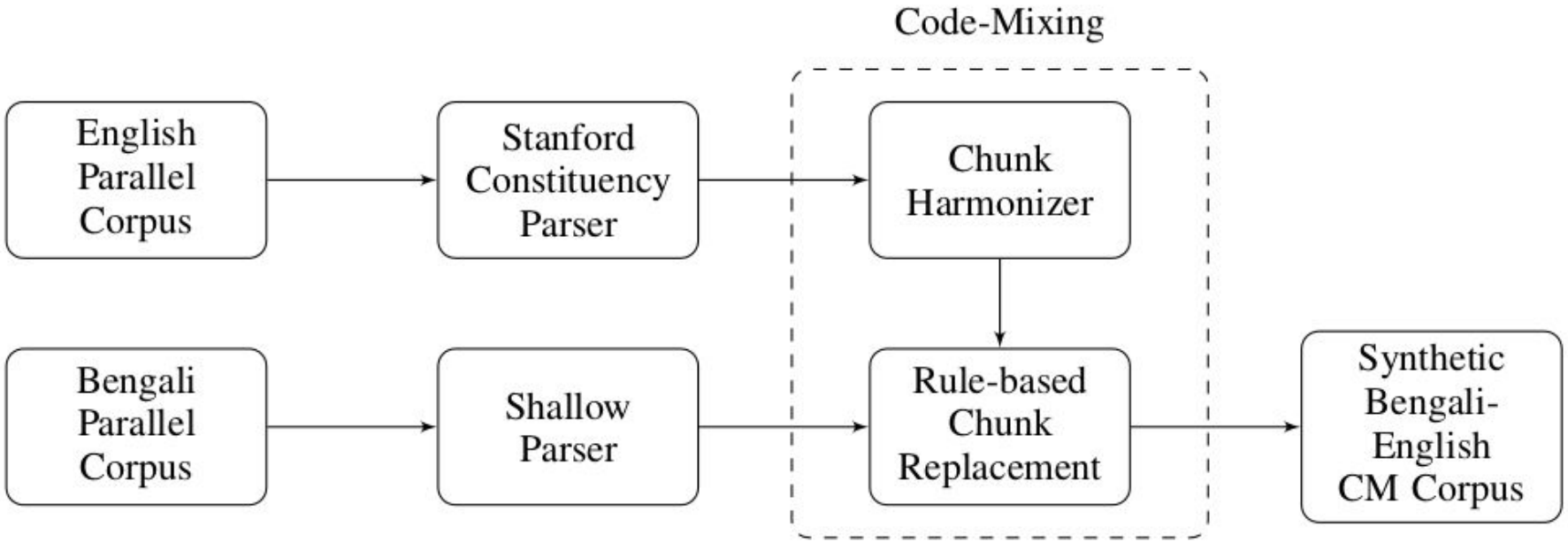
- 500 Bengali-English tweets from Twitter
- code-mixing ratio of 30:70(%)
- Universal Dependency Annotations

$$\frac{1}{n}\sum_{s=1}^{n}\frac{E_s}{M_s + E_s}$$

E_s, = embedded
M_s = matrix



Code-Mixing Data Synthesis



Code-Mixing Process

Chunk Harmonizer

1. Separate the *coordinating conjunction*
2. Combine the *adverbs of degree* with preceding NP
3. Convert PP to NP, separate from VP
4. Split NP at genitives

Rule-based Chunk Replacement

- Closed Class Constraint (Sridhar and Sridhar, 1980; Joshi, 1982)
- Replace Bengali NP and JJP with English
- Retain Bengali Post positions

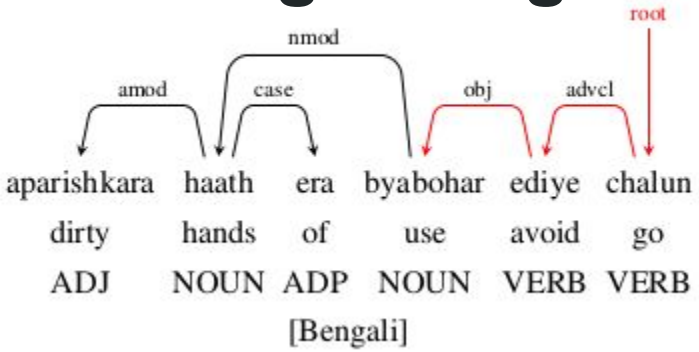
(NP Your self-confidence) (ADVP also) (VP increases (PP with (NP teeth))) ENGLISH

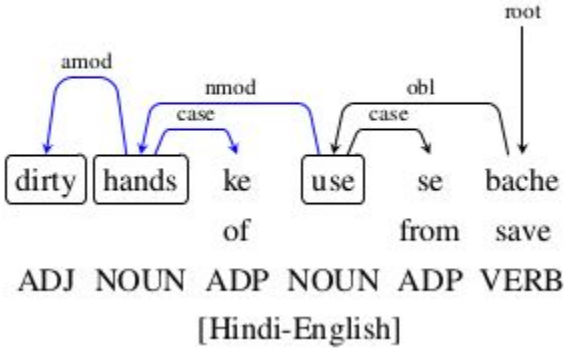
(NP daanter “teeth” jonyo “for”) (NP aapnaar “your”) (NP aatmaviswas “self-confidence” o“also”) (VP baadhe “increases”) BENGALI

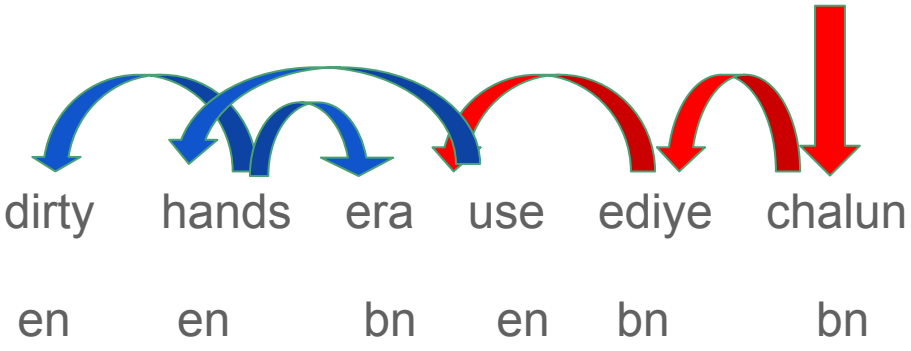
(NP Your) (NP self-confidence also) (VP increases) (NP with teeth) HARMONIZED ENGLISH

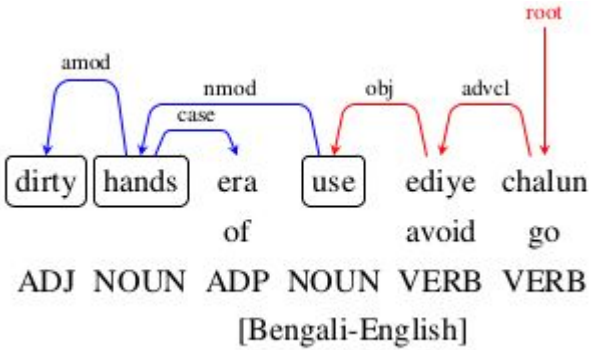
(NP teeth er “of” jonyo “for”) (NP aapnaar “your”) (NP self-confidence also) (VP baadhe “increases”) BENGALI -ENGLISH CM

Synthetic Bengali-English Treebank



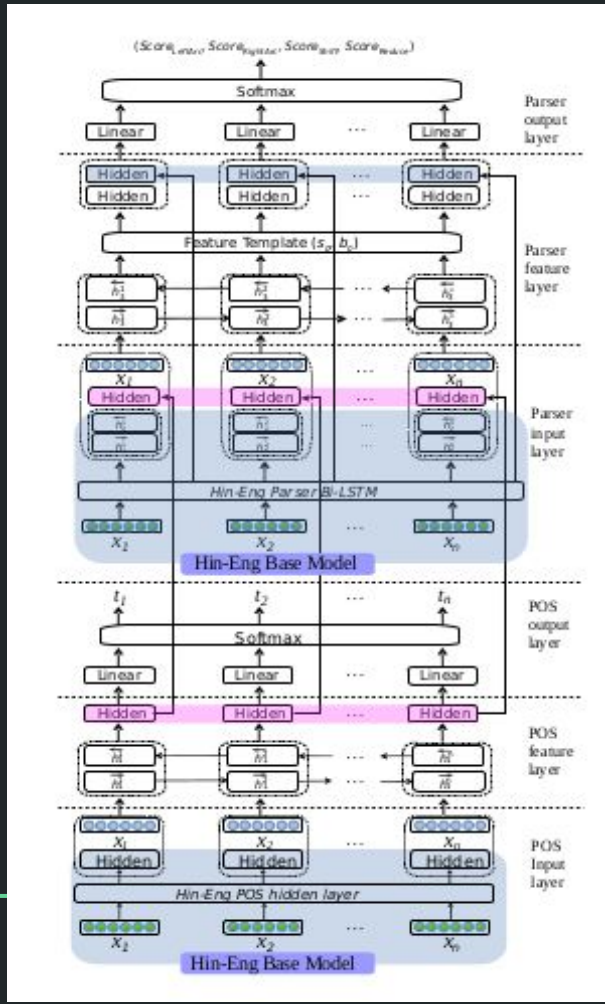






Neural-Stack based Dependency Parser

- Bhat et al. (2018) for Hindi-English
- transition-based parser (Kiperwasser and Goldberg, 2016)
- Joint learning of POS and Parsing (Zhang and Weiss, 2016; Chen et al., 2016)
- enhanced by neural stacks to incorporate monolingual syntactic knowledge with the CM model



Experiments and Results

Bilingual + Gold BE

POS	UAS	LAS
79.39	62.78	49.38

- Small CM Training Data Size (140)
- Utilizes English(12k), Bengai Treebank (9k)
- Not enough CM grammer

Trilingual
+ Gold (BE +HE)

POS	UAS	LAS
87.43	74.42	60.04

- + Utilizes existing BE(140), HE data (1448) CM data
- + Utilizes English(12k), Bengai Treebank (9k), Hindi Treebank (11k)

(Trilingual + Syn BE)
+ Gold (BE+HE)

POS	UAS	LAS
89.63	76.24	61.41

- + Utilizes Syn-BE (3643)
- + Utilizes existing BE(140), HE data (1448) CM data
- + Utilizes English(12k), Bengai Treebank (9k), Hindi Treebank (11k)

Conclusion

Limitations

1. Error Propagation as automatically annotated
2. Not all cases of code-mixing is covered

Contribution

1. State of the art POS tagger + Dependency Parser for Bengali English CM (**89.63 76.24 61.41**)
2. 500 Bengali-English UD annotated tweets
3. Synthetic-BE Data to help in other NLP CM systems

Thank You!