

Pratik Mehta

Machine Learning
Deep Learning
NLP

MS UMass CS, 2016-2018

Seeking: Internship: Summer 2017

Research Engineer, CFILT - IIT Bombay

Research Intern, CFILT - IIT Bombay

BE CS, University of Mumbai, Machine Translation

Overcoming data sparsity in Statistical Machine Translation for morphologically rich languages

Problem

- For many Indian languages, larger parallel corpora to train SMT systems are hard to find

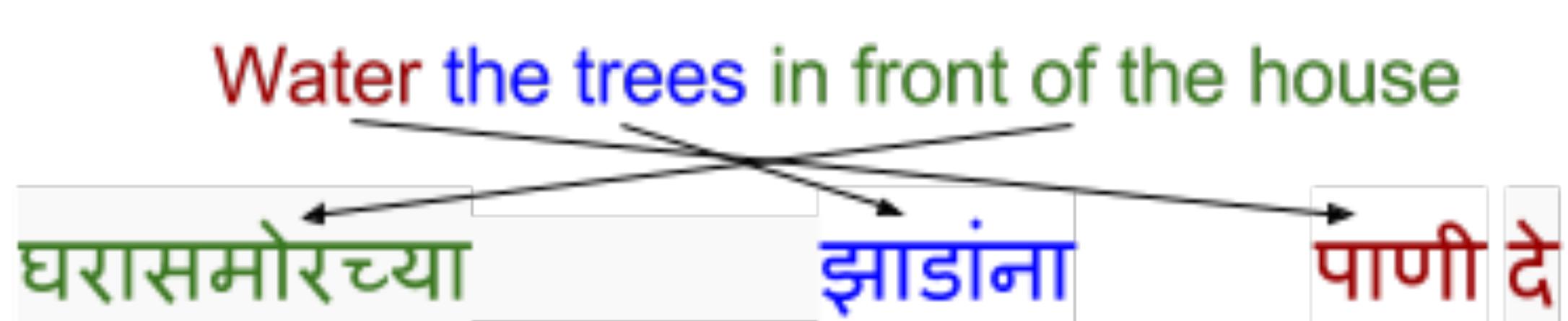
English: Water the trees in front of the house

Marathi: घरासमोरच्या झाडांना पाणी दे.

1. Agglutination: झाडांना
(to the trees)

2. Morph complexity: घर- ा-समोर-च्या
(in front of the house)

3. Structural divergence (Word order):



- Agglutination causes decreased occurrence counts of content words in the corpus, leading to sharp drops in translation accuracy
- A rich morpheme set in the target language compounds the issue

Current Work

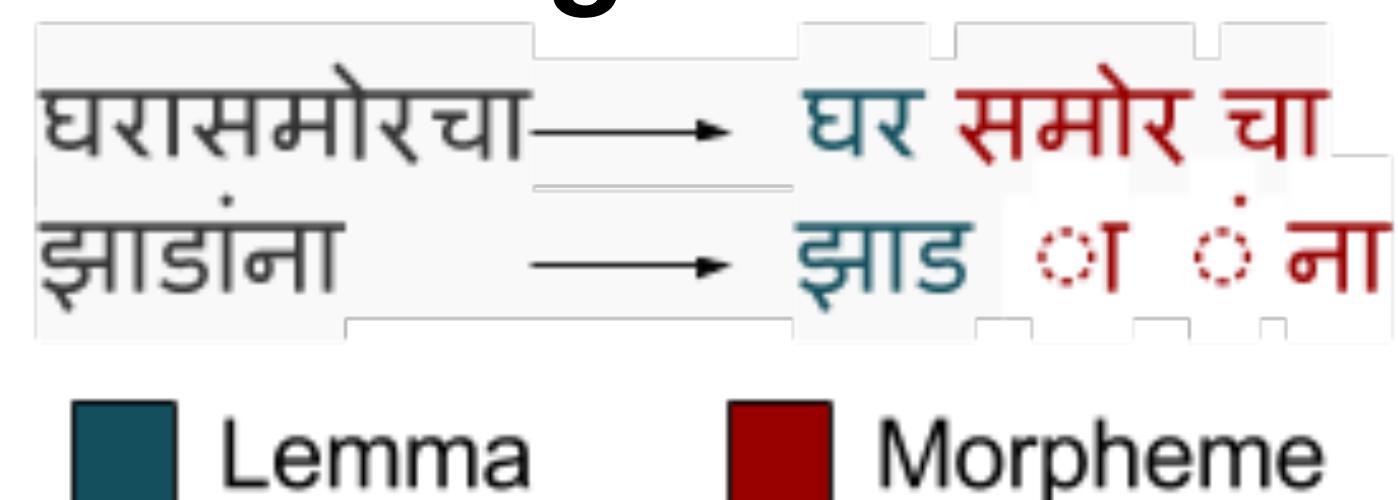
Course-work: Machine Learning (689)
Deep Learning (697L)

The AI history project: A review of foundational papers set forth in AI, information theory and deep learning by Von Neumann, Minsky, Shannon, Rosenblatt and Hinton

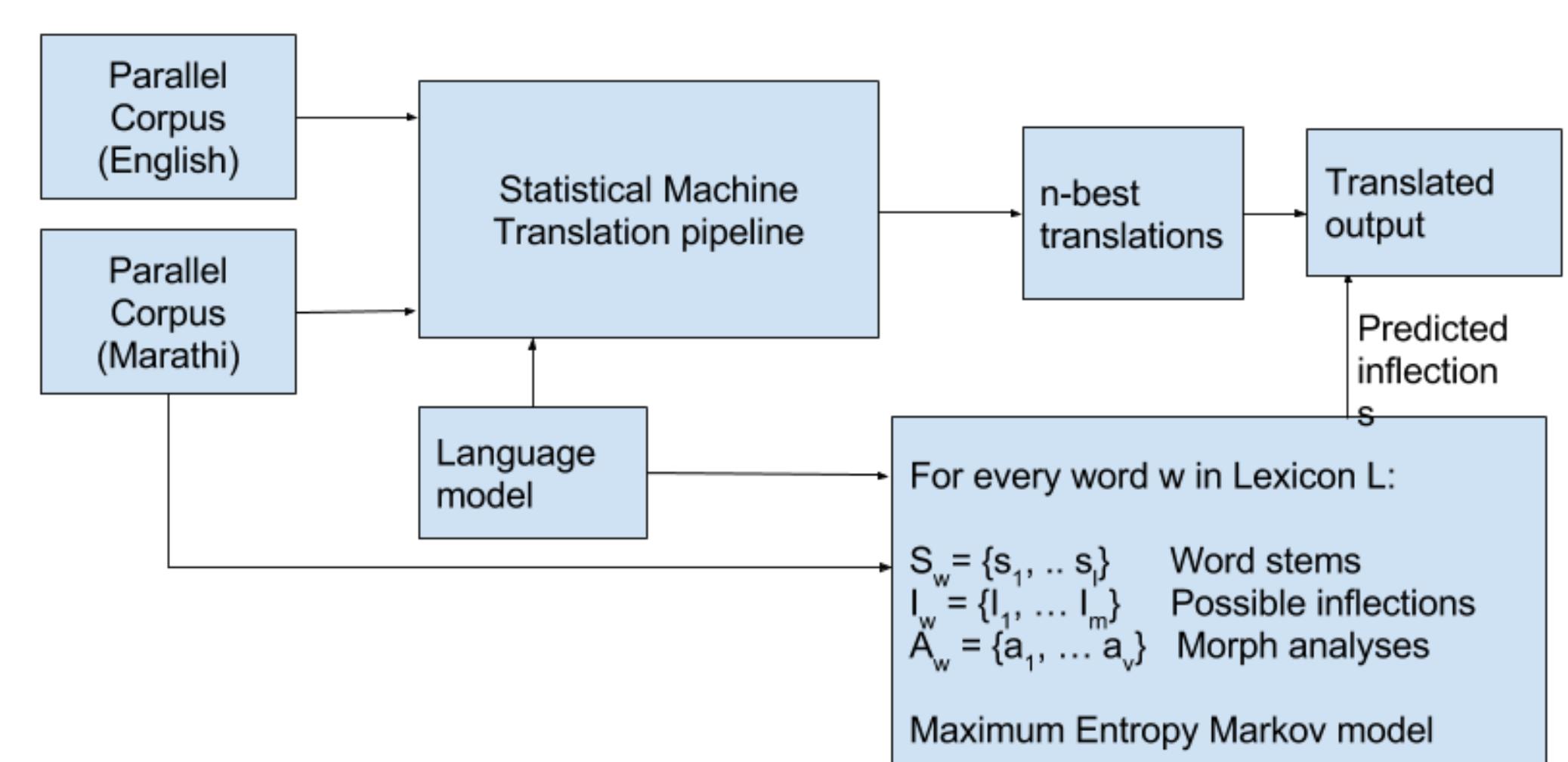
Seeking...
Internship, Summer 2017

Solutions

Unsupervised morphological segmentation



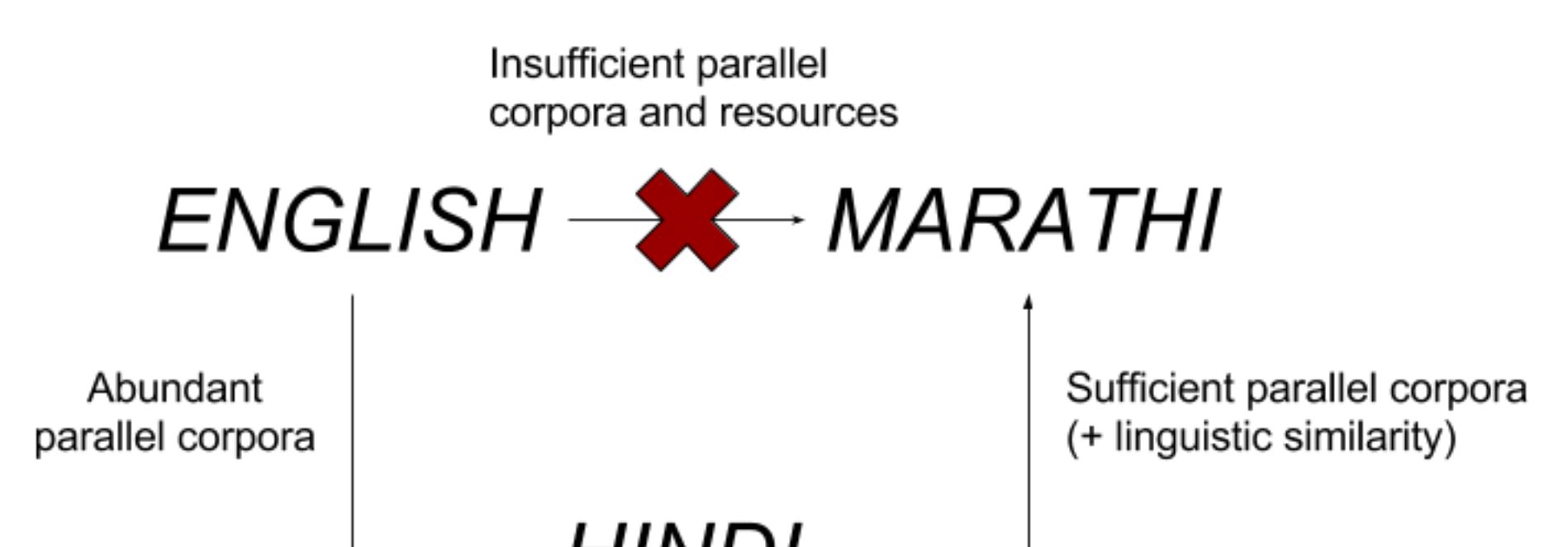
Morphology generation models



Rule-based source pre-ordering

Subject - Verb - Object
We will make America great again
↓
We America again great will make
Subject - Object - Verb

Pivot-based SMT



Results

Improvement in BLEU scores:
Segmentation: +1.9
Source-preordering: +1.39