

## Corpus Paths (relative to src/)

train.en

data/train.en

train.hi

data/train.hi

## Training

Max training pairs

500000



IBM1 EM iterations

10

Phrase extraction pairs

300000



Max phrase length

5

Random seed

42



## Decoding

Beam size

20

Candidates/phrase

10

# EN→HI Statistical Machine Translation (SMT)

IBM Model 1 word alignment → phrase extraction → phrase table →  
bigram Hindi LM → noisy-channel decoding

## Step 1 — Load a saved model (fast) OR Train a new model

Load Model

Train Model

Save Model

## System Requirements Output

| Sentence pa... | EN vocab | HI vocab | HI LM vocab |
|----------------|----------|----------|-------------|
| 50...          | 87...    | 12...    | 12...       |

IBM iters: 10 | Phrase pairs: 300000 | Max phrase len: 5 | Build sec: 2496.56

## Top learned lexical translation probabilities: $t(hi|en)$

the → के (0.130), की (0.096), को (0.086), है (0.085), में (0.063),  
और (0.062), का (0.060), , (0.043)

and → और (0.648), ( (0.037), ) (0.036), की (0.024), भी (0.020),  
को (0.019), से (0.018), है (0.017)

**of** → की (0.202), के (0.169), का (0.132), से (0.091), में (0.046),  
और (0.040), है (0.029), ( (0.026)

**to** → को (0.136), के (0.118), लिए (0.086), में (0.068), की (0.055),  
कि (0.049), है (0.048), करने (0.048)

**is** → है (0.565), है। (0.074), वह (0.026), ( (0.023), ) (0.023), और  
(0.023), तो (0.020), यह (0.020)

**in** → में (0.594), के (0.048), है (0.033), और (0.033), की (0.033),  
पर (0.026), से (0.022), जो (0.017)

**a** → एक (0.311), के (0.090), है (0.068), की (0.043), से (0.039), का  
(0.033), को (0.031), कोई (0.029)

**you** → तुम (0.399), हो (0.104), तो (0.052), आप (0.045), (  
(0.037), ) (0.037), कि (0.029), से (0.026)

---

## Translate (Noisy Channel Decoding)

English input

One prince charming

Translate

Translation completed.

**HYP:** एक प्रिंस charming

TM log: -28.130 | LM log: -40.683 | Total: -76.950

