

Machine Translation and its modeling with implementation

Presented by
Debajyoty Banik
IIT Patna

For any clarification feel free to contact at
debajyoty.banik@gmail.com

06/12/2017



Machine Translation

- Automatically translate one natural language into another

I love India



मैं भारत से प्यार करता हूँ

Type of MT

- Rule Based MT
- SMT
- NMT
- Hybrid MT

For Translation Ambiguity Resolution is Required

- For correct translation semantic and syntactic ambiguities must be resolved properly:
 - “John **plays** the guitar.” → “जॉन गिटार **बजाता** है।”
 - “John **plays** football.” → “जॉन फुटबॉल **खेलता** है।”
- Some of statement does not have direct meaning.
 - “Out of sight, out of mind.” → “Invisible idiot.”
- Semantic role ambiguity
 - मुझे आपको मिठाई खिलानी पड़ेगी



Parallel corpus

English	Hindi
Fresh breath and shining teeth enhance your personality.	ताजा साँसें और चमचमाते दाँत आपके व्यक्तित्व को नखिरते हैं।
Your self-confidence also increases with teeth.	दाँतों से आपका आत्मविश्वास भी बढ़ता है।
Bacteria stay between our gums and teeth.	हमारे मसूढ़ों और दाँतों के बीच बैक्टीरिया मौजूद होते हैं।

How to Build an SMT System

- Start with a large parallel corpus
 - Consists of document pairs (document and its translation)
- Sentence alignment: in each document pair automatically find those sentences which are translations of one another
 - Results in sentence pairs (sentence and its translation)
- Word alignment: in each sentence pair automatically annotate those words which are translations of one another
 - Results in word-aligned sentence pairs
- Automatically estimate a statistical model from the word-aligned sentence pairs
 - Results in model parameters
- Given new text to translate, apply model to get most probable translation

Preprocessing

- Tokenization
- Truecasing
- Cleaning
- etc

How to Build an SMT System

- Construct a function g which, given a sentence in the source language and a hypothesized translation into the target language, assigns a goodness score
 - $g(\text{I love India, मैं भारत से प्यार करता हूँ}) = \text{high number}$
 - $g(\text{I love India, मैं हरे पेड़ से प्यार करता हूँ}) = \text{low number}$

SMT modeling

- We wish to build a machine translation system which given a Foreign sentence (English) “f” produces its Hindi translation “e”
 - We build a model of $P(e | f)$, the probability of the sentence “e” given the sentence “f”
 - To translate a Foreign text “f”, choose the Hindi text “e” which maximizes $P(e | f)$

SMT Model

$$e_{best} = \arg \max_e P(e | f)$$

$$= \arg \max_e \left[\prod_{i=1}^I \{ \Phi(\bar{f}_i | \bar{e}_i)^{\lambda_\phi} d(start_i - end_{i-1} - 1)^{\lambda_d} \} P_{LM}(e)^{\lambda_{LM}} \right]$$

Translation Model

- Translation model probabilities have to be computed from parallel corpora.

$$P(f | e)$$

Language Model (LM)

$$\log(P_{LM}(C^s)) = \sum_{j=1}^{|C^s|} \log P(c_j^s \mid c_1^s, c_2^s, \dots, c_{j-1}^s)$$

Example of LM: IRSTLM


```
mkdir lm
cd lm
~/smt/irstlm/bin/add-start-end.sh <~/set_eng_to_hindi/corpora_train.clean.hi
>corpora_train.lm.sh.hi

export IRSTLM=$HOME/smt/irstlm

~/smt/irstlm/bin/build-lm.sh -i ~/set_eng_to_hindi/lm/corpora_train.lm.sh.hi -t ./temp -p
-s improved-kneser-ney -n 3 -o ~/set_eng_to_hindi/lm/corpora_train.lm.final.hi

~/smt/irstlm/bin/compile-lm --text=yes corpora_train.lm.final.hi.gz
corpora_train.lm.final.arpa.hi
```

Training for SMT model



```
~/smt/mosesdecoder/scripts/training/train-model.perl  
-root-dir train -corpus ~/set_eng_to_hindi/corpora_train.cleaned  
-f en -e hi -alignment grow-diag-final-and -reordering  
msd-bidirectional-fe -lm 0:3:$HOME/set_eng_to_hindi/lm/  
corpora_train.lm.arpa.binary:1  
-external-bin-dir ~/smt/bin | tee training.out
```

Fixing λ Values: Tuning

$$\lambda_{\Phi|d|LM} (new) = \lambda_{\Phi|d|LM} (old) + f(errors)$$

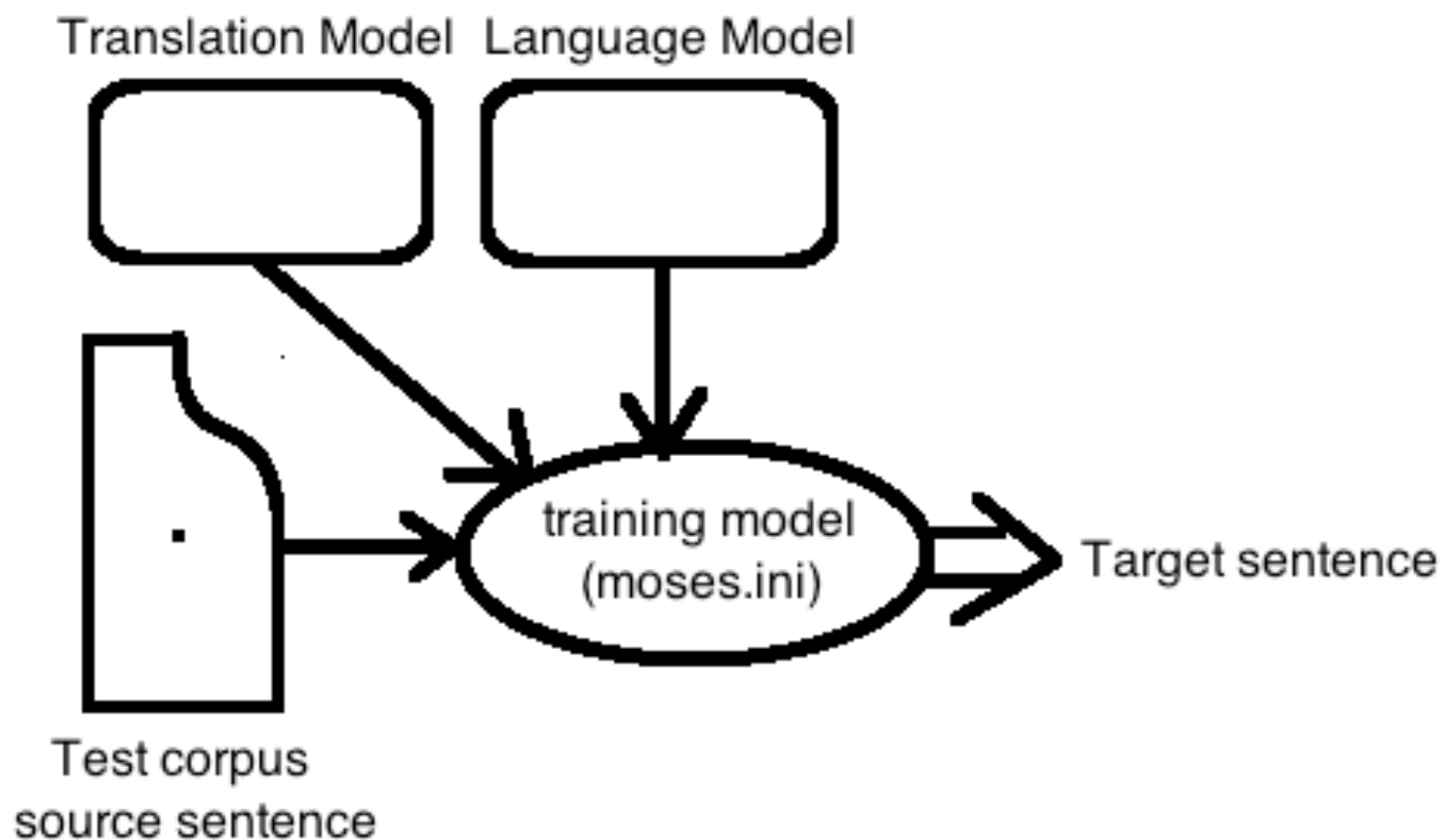


MT parameter estimation

```
~/smt/mosesdecoder/scripts/training/mert-moses.pl tune/corpora_tune.clean.en tune/  
corpora_tune.clean.hi ~/smt/mosesdecoder/bin/moses working/train/model/moses.ini --mertdir  
~/smt/mosesdecoder/bin
```



Decoding



Decoding



```
~/smt/mosesdecoder/bin/moses -f ~/set_eng_to_hindi/mert-work/moses.ini <test.true.en  
>test.out.hi.text
```

Dataset

	Sentences		Token	
	English	Hindi	English	Hindi
Training Set	5000	5000	87233	92507
Development set 1	39	39	8879	9033
Development set 2	500	500	8879	9033
Test set	100	100	1772	1763

MT Evaluation

BLEU

METEOR

LEBLEU

NIST

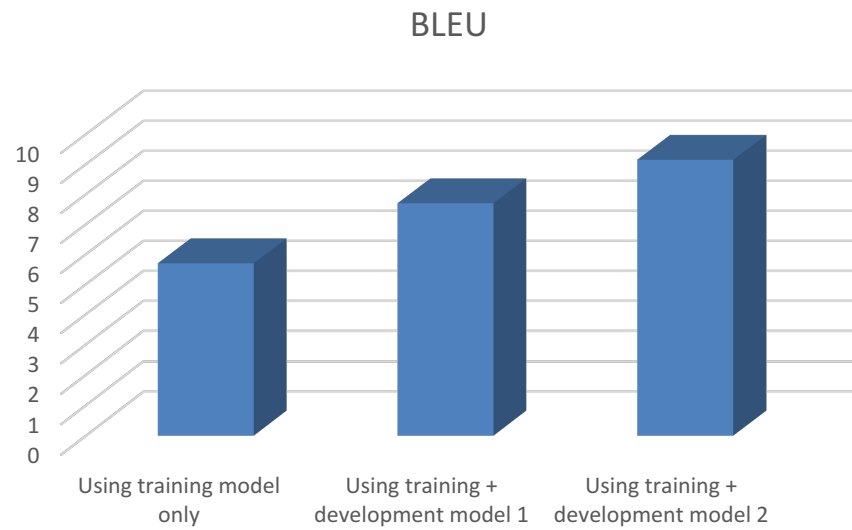
WER

etc



Results

Results





Success
is dependent on effort.

ALL THE BEST