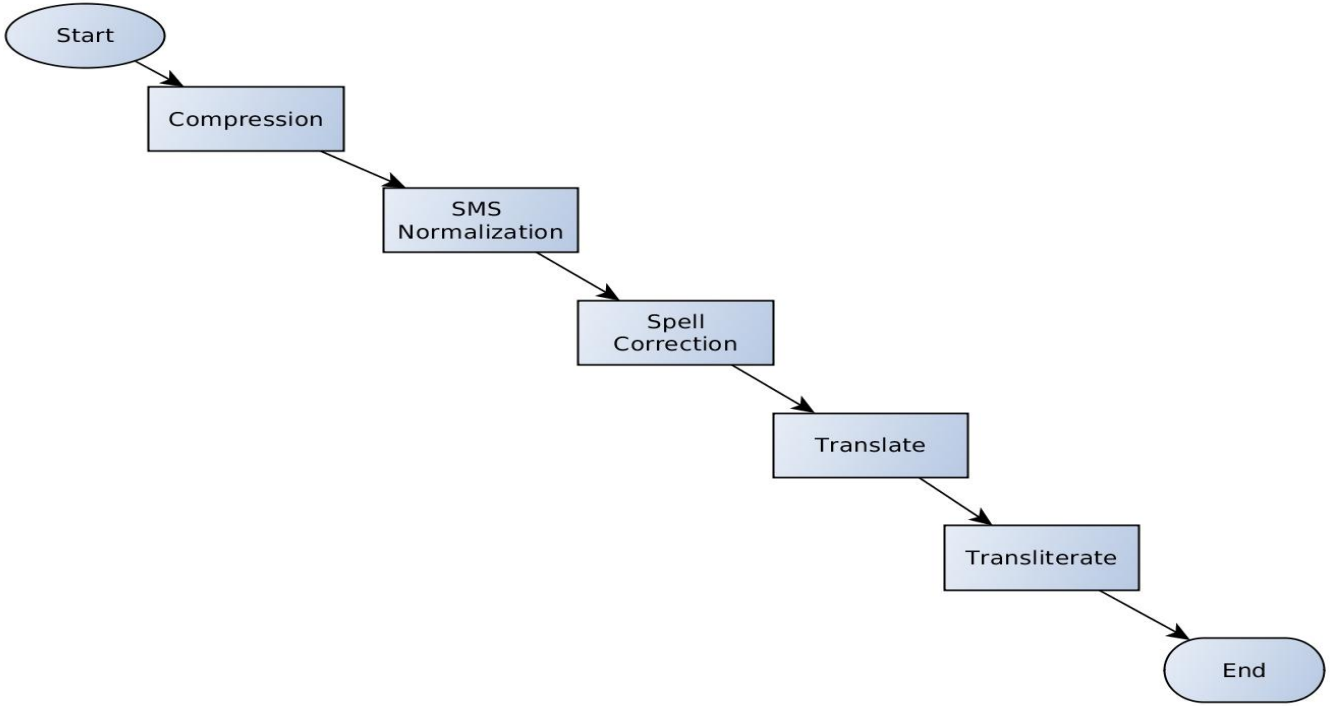
SMS Translator

By
Aman Madaan (13305004)
Naman Gupta (133050012)

CS 712, Spring 2014, IIT Bombay

Big Picture


Start
Compression
SMS Normalization
Spell Correction
Translate
Transliterate
End

Requirements Fulfilled

Core Requirements:

1. Mobile interacting with the server directly/indirectly should be in place.

   Translation Server :10.129.28.112 : 1245

   Normalisation Server : 10.129.26.85 : 1246

2. Text normalization (high quality)
3. Translation should happen (80% of words should be translated)
4. Translation should be visible on the device (cell phone, etc.)
5. Literature survey for normalization

Requirements Fulfilled

1. Translation with adequacy score >= 3

We add transliteration as the last step. This means that there never be any English words in the output, user can always read

2. Translation with fluency score >=3

Translations seem to be relatively satisfactory!

Innovative Ideas

- Compression
- Translation Caching
- Scoring
- Transliteration for untranslated words : no part of foreign Language in the output

Innovative Ideas

- Application as Learning Tool
  - Text to Speech
  - Visualisation of Translation Process.

- Normalization : Statistical + Rule Base

- Comparison with Heuristic, Dictionary Based methods.

Compression

Compression : Intuition

- There can never be three consecutive letters in English!
- Smash all the repeat windows of length greater than >2 to length 2
- Eg : pleeeeeeeeaaaaaaassseeeee -> pleeaassee
- Two choices for each window, O(2^n), start with vowels
- n is typically small , can brute force (Avg 5.1 characters per word)

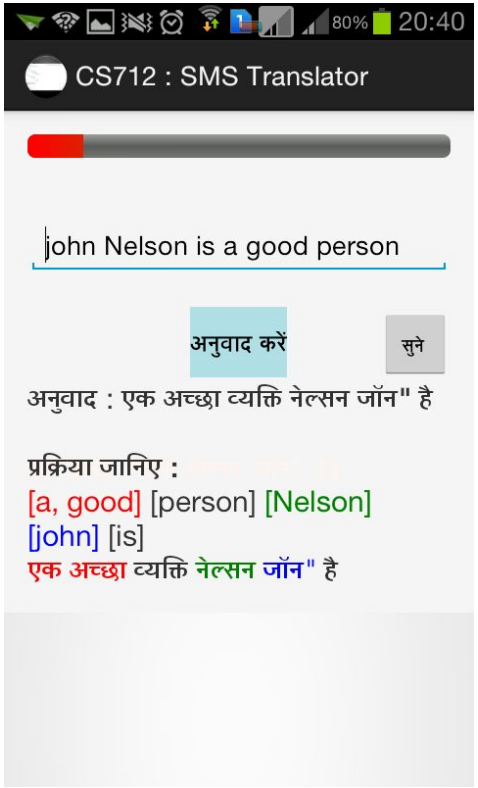
Compression

```
String findFirstMatch(String s, int currPos, int adj) {
    //System.out.println(s + ", " + currPos);
    if(dict.contains(s)) {
        return s;
    }
    if(currPos == repeatPosition.size()) {
        return null;
    }
    String res = findFirstMatch(s, currPos + 1, adj);
    if(res != null) {
        return res;
    }
    int repeatsAt = repeatPosition.get(currPos) - adj;
    String temp;
    if(repeatsAt == s.length() - 2) { //second last char
        temp = s.substring(0, s.length() - 1);
    }else {
        System.out.println(s);
        temp = s.substring(0, repeatsAt + 1) + s.substring(repeatsAt + 2,
    }
    res = findFirstMatch(temp, currPos + 1, adj + 1);
    return res;
}
```

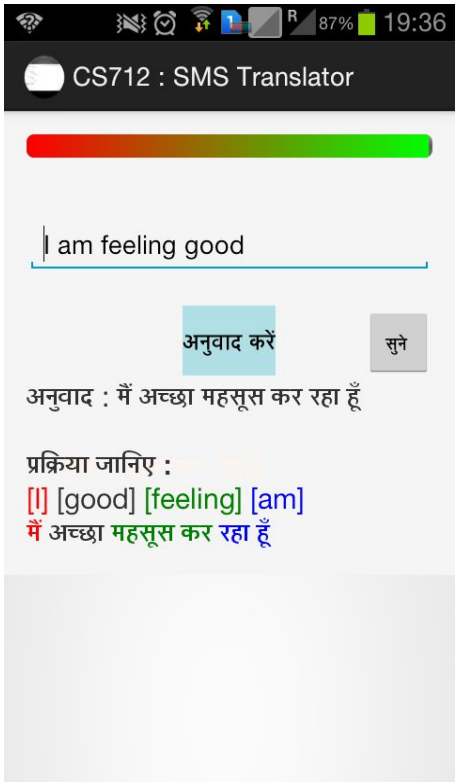
heell, 0, 0
heell, 1, 0
heell, 2, 0
hell, 2, 1
hell

Scoring

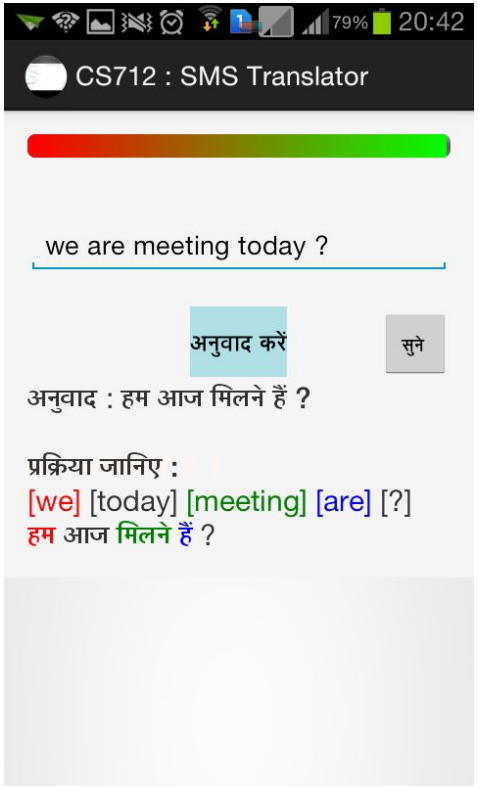
Scoring

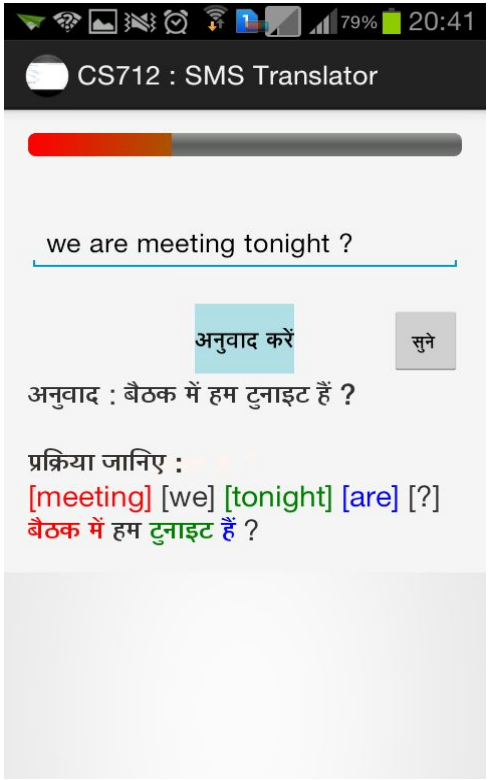


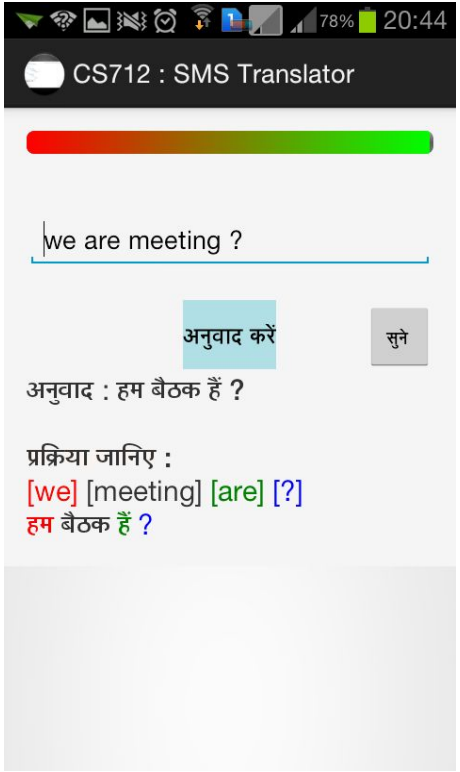
- Translation score given by moses is used after suitable transformation
- Score is in log space, not considered
- Score is penalized for untranslatable words (names etc.)



Scoring


79% 20:42
CS712 : SMS Translator
we are meeting today ?
अनुवाद करें
सुने
अनुवाद : हम आज मिलने हैं ?
प्रक्रिया जानिए :
[we] [today] [meeting] [are] [?]
हम आज मिलने हैं ?


79% 20:41
CS712 : SMS Translator
we are meeting tonight ?
अनुवाद करें
सुने
अनुवाद : बैठक में हम टुनाइट हैं ?
प्रक्रिया जानिए :
[meeting] [we] [tonight] [are] [?]
बैठक में हम टुनाइट हैं ?


78% 20:44
CS712 : SMS Translator
we are meeting ?
अनुवाद करें
सुने
अनुवाद : हम बैठक हैं ?
प्रक्रिया जानिए :
[we] [meeting] [are] [?]
हम बैठक हैं ?

Learning Tool

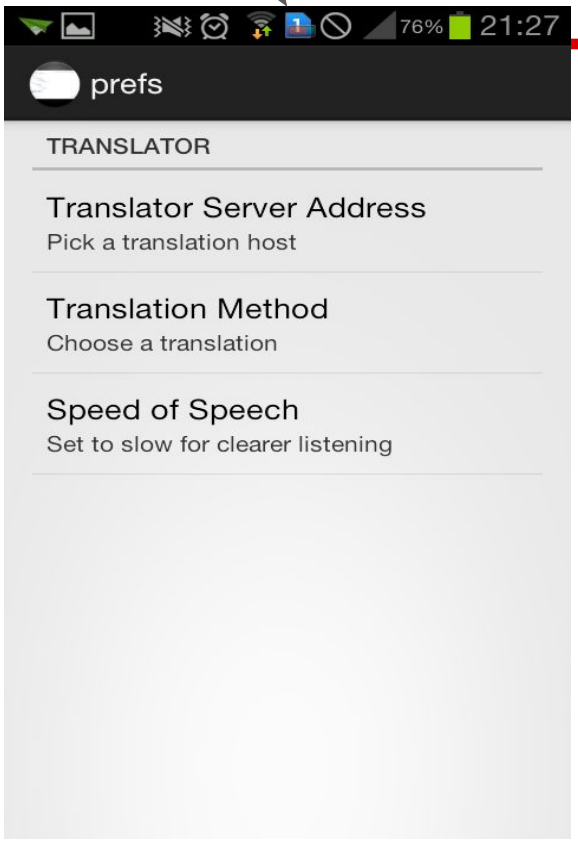
Translator as a learning tool

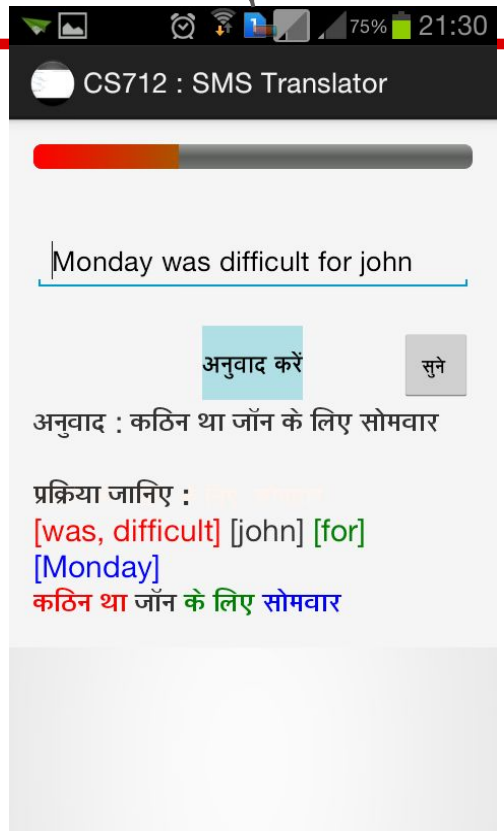
- Provision for saving frequent translations of good quality
- Can listen to the foreign language sentence
- Visualization of the translation process

TTS

Translation Process

Caching





Normalisation

SMS Text Normalization

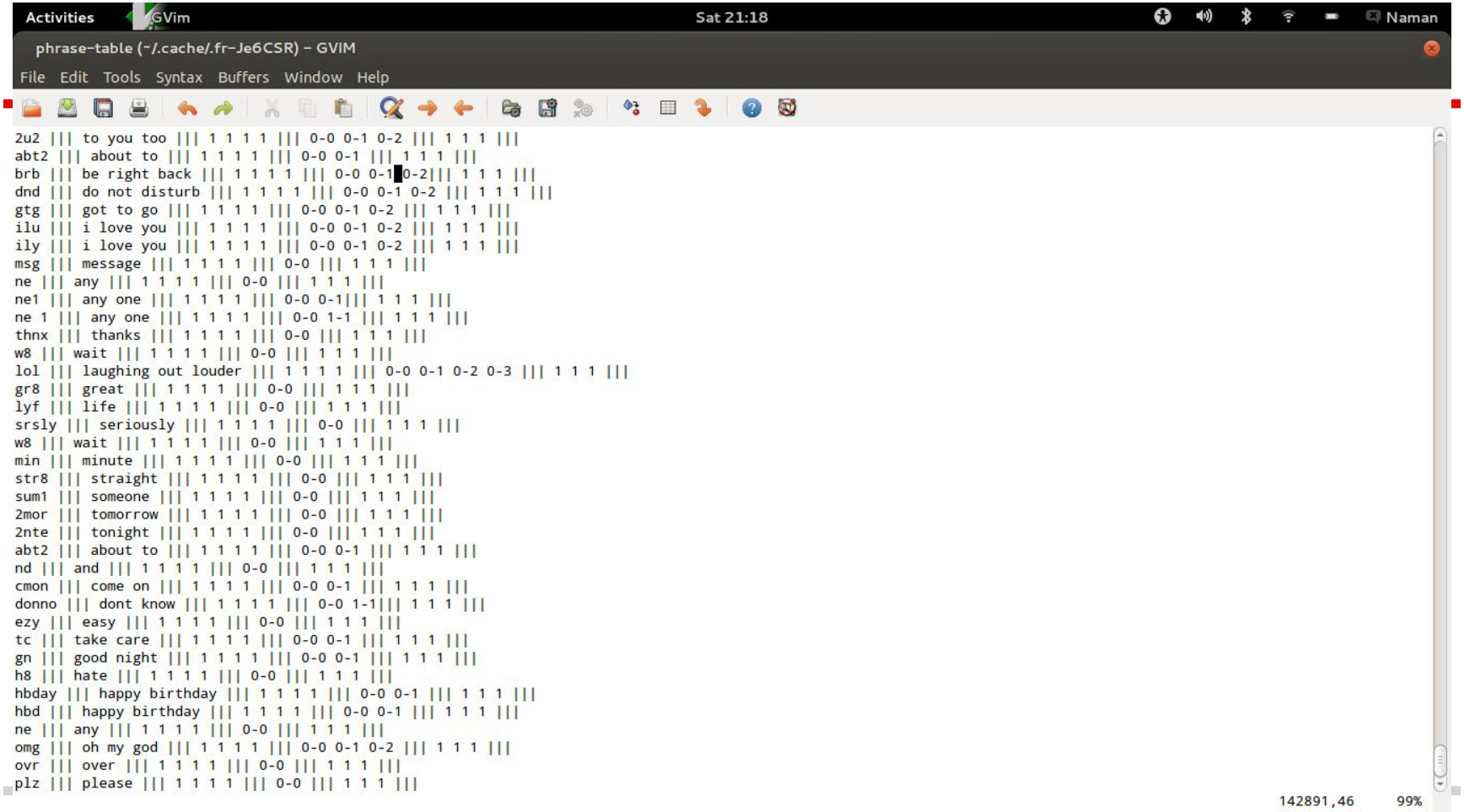
- Statistical + Rule Based

- Statistical
  - Data from Stanford SMS text Normalisation
  - Language Model : 15000
  - Training : 2470
  - Server running at 10.129.26.85 : 1245
    - Investigating SMS Text Normalization using Statistical Machine Translation by Karthik Raghunathan, Stefan Krawczyk

SMS Text Normalization

- Rule Based
  - Updated Phrase table with 111 pairs of unnormalized-normalized words

- Snapshot on next Slide


Activities
GVim
Sat 21:18
Naman
phrase-table (~/.cache/.fr-Je6CSR) - GVIM
File Edit Tools Syntax Buffers Window Help
2u2 ||| to you too ||| 1 1 1 1 ||| 0-0 0-1 0-2 ||| 1 1 1 |||
abt2 ||| about to ||| 1 1 1 1 ||| 0-0 0-1 ||| 1 1 1 |||
brb ||| be right back ||| 1 1 1 1 ||| 0-0 0-1 0-2||| 1 1 1 |||
dnd ||| do not disturb ||| 1 1 1 1 ||| 0-0 0-1 0-2 ||| 1 1 1 |||
gtg ||| got to go ||| 1 1 1 1 ||| 0-0 0-1 0-2 ||| 1 1 1 |||
ilu ||| i love you ||| 1 1 1 1 ||| 0-0 0-1 0-2 ||| 1 1 1 |||
ily ||| i love you ||| 1 1 1 1 ||| 0-0 0-1 0-2 ||| 1 1 1 |||
msg ||| message ||| 1 1 1 1 ||| 0-0 ||| 1 1 1 |||
ne ||| any ||| 1 1 1 1 ||| 0-0 ||| 1 1 1 |||
ne1 ||| any one ||| 1 1 1 1 ||| 0-0 0-1||| 1 1 1 |||
ne 1 ||| any one ||| 1 1 1 1 ||| 0-0 1-1 ||| 1 1 1 |||
thnx ||| thanks ||| 1 1 1 1 ||| 0-0 ||| 1 1 1 |||
w8 ||| wait ||| 1 1 1 1 ||| 0-0 ||| 1 1 1 |||
lol ||| laughing out louder ||| 1 1 1 1 ||| 0-0 0-1 0-2 0-3 ||| 1 1 1 |||
gr8 ||| great ||| 1 1 1 1 ||| 0-0 ||| 1 1 1 |||
lyf ||| life ||| 1 1 1 1 ||| 0-0 ||| 1 1 1 |||
srsly ||| seriously ||| 1 1 1 1 ||| 0-0 ||| 1 1 1 |||
w8 ||| wait ||| 1 1 1 1 ||| 0-0 ||| 1 1 1 |||
min ||| minute ||| 1 1 1 1 ||| 0-0 ||| 1 1 1 |||
str8 ||| straight ||| 1 1 1 1 ||| 0-0 ||| 1 1 1 |||
sum1 ||| someone ||| 1 1 1 1 ||| 0-0 ||| 1 1 1 |||
2mor ||| tomorrow ||| 1 1 1 1 ||| 0-0 ||| 1 1 1 |||
2nte ||| tonight ||| 1 1 1 1 ||| 0-0 ||| 1 1 1 |||
abt2 ||| about to ||| 1 1 1 1 ||| 0-0 0-1 ||| 1 1 1 |||
nd ||| and ||| 1 1 1 1 ||| 0-0 ||| 1 1 1 |||
cmon ||| come on ||| 1 1 1 1 ||| 0-0 0-1 ||| 1 1 1 |||
donno ||| dont know ||| 1 1 1 1 ||| 0-0 1-1||| 1 1 1 |||
ezy ||| easy ||| 1 1 1 1 ||| 0-0 ||| 1 1 1 |||
tc ||| take care ||| 1 1 1 1 ||| 0-0 0-1 ||| 1 1 1 |||
gn ||| good night ||| 1 1 1 1 ||| 0-0 0-1 ||| 1 1 1 |||
h8 ||| hate ||| 1 1 1 1 ||| 0-0 ||| 1 1 1 |||
hbday ||| happy birthday ||| 1 1 1 1 ||| 0-0 0-1 ||| 1 1 1 |||
hbd ||| happy birthday ||| 1 1 1 1 ||| 0-0 0-1 ||| 1 1 1 |||
ne ||| any ||| 1 1 1 1 ||| 0-0 ||| 1 1 1 |||
omg ||| oh my god ||| 1 1 1 1 ||| 0-0 0-1 0-2 ||| 1 1 1 |||
ovr ||| over ||| 1 1 1 1 ||| 0-0 ||| 1 1 1 |||
plz ||| please ||| 1 1 1 1 ||| 0-0 ||| 1 1 1 |||
142891,46
99%

Techniques Explored

- Dictionary Based Translations
- Alignment Heuristic Based
- Moses Based*

**These slides*

Evaluation

- We compare our translator (ANT) with
  - Sata Anuvadak(SATA) from CFILT, IIT Bombay and
  - Google Translator(Google)

- The input to the google translator and sata anuvadak was normalized.

Evaluation

Input	Actual norm	Our Norm	ANT	SATA*	Google*
dat s a goooood question	That is a good question	that is a good question	एक अच्छा सवाल है कि	एक अच्छी बात है कि है	यह एक अच्छा सवाल है
plz inform ur sis abt dis	Please inform your sister about this	Please inform your sister about this	इस बारे में अपनी बहन सूचित करें चलोगे	इस बारे में अपनी बहन कृपया Inform	इस बारे में अपनी बहन को सूचित करें
I wl try dat l8r	I will try that later	I will try that later	कोशिश करें कि बाद में मैं	कोशिश करें कि आगे चलकर मैं होगा	मुझे लगता है कि बाद में कोशिश करेंगे
cn u say it agn plz	can you say it again please	can you say it again please	कहते हैं यह सकता है फिर आप चलोगे	आप वह पुनः please कह सकते हैं	आप इसे फिर से कृपया कह सकते हैं

Evaluation

Input	Actual norm	Our Norm	ANT	SATA*	Google*
nethng new gng on	Anything new going on	Anything new going on	नया कुछ भी पर जाने वाले	नए जाने पर Anything	पर जा रहे कुछ नया
tell me more abt it	tell me more about it	tell me more about it	इसके बारे में अधिक मुझे बता	मुझे यह अधिक के बारे में बताते हैं	इसके बारे में मुझे और बताएँ
I m feeeling gooood	I am feeling good	I am feeling good	मैं अच्छा महसूस कर रहा हूँ	मुझे अच्छा महसूस कर रही हूँ	मैं अच्छा महसूस कर रहा हूँ
shd v go 2 ur house den	should we go to your house then	should we go to your house then	हम अपने घर के करना चाहिए जाएँ तो	हम घर पर जाएँ तो	हम तो अपने घर के पास जाना चाहिए

Evaluation

Input	Actual norm	Our Norm	ANT	SATA*	Google*
Bill Clinton had a meeting today	Anything new going on	Anything new going on	बिल क्लिंटन में एक बैठक था आज	आज संगम Clinton Bill थी	बिल क्लिंटन आज एक बैठक की थी
m soooooooooo haaaaappppppp yyyyyyyyy John	am so happy john	am so happy john	जॉन बड़ा खुश हूं	इतना john रहा है	बहुत खुश जॉन हूँ
cn u plz chk nw	can you please check now	can you please check now	अब आप सकता है चलोगे जाँच	अब आप जाँच please सकते हैं	आप अब जाँच कृपया कर सकते हैं
Monday was difficult for john	Monday was difficult for john	Monday was difficult for john	कठिन था जॉन के लिए सोमवार	सोमवार कठिन john थी	सोमवार जॉन के लिए मुश्किल था

THANKS