TransChat: A Cross Lingual Indian Language IM

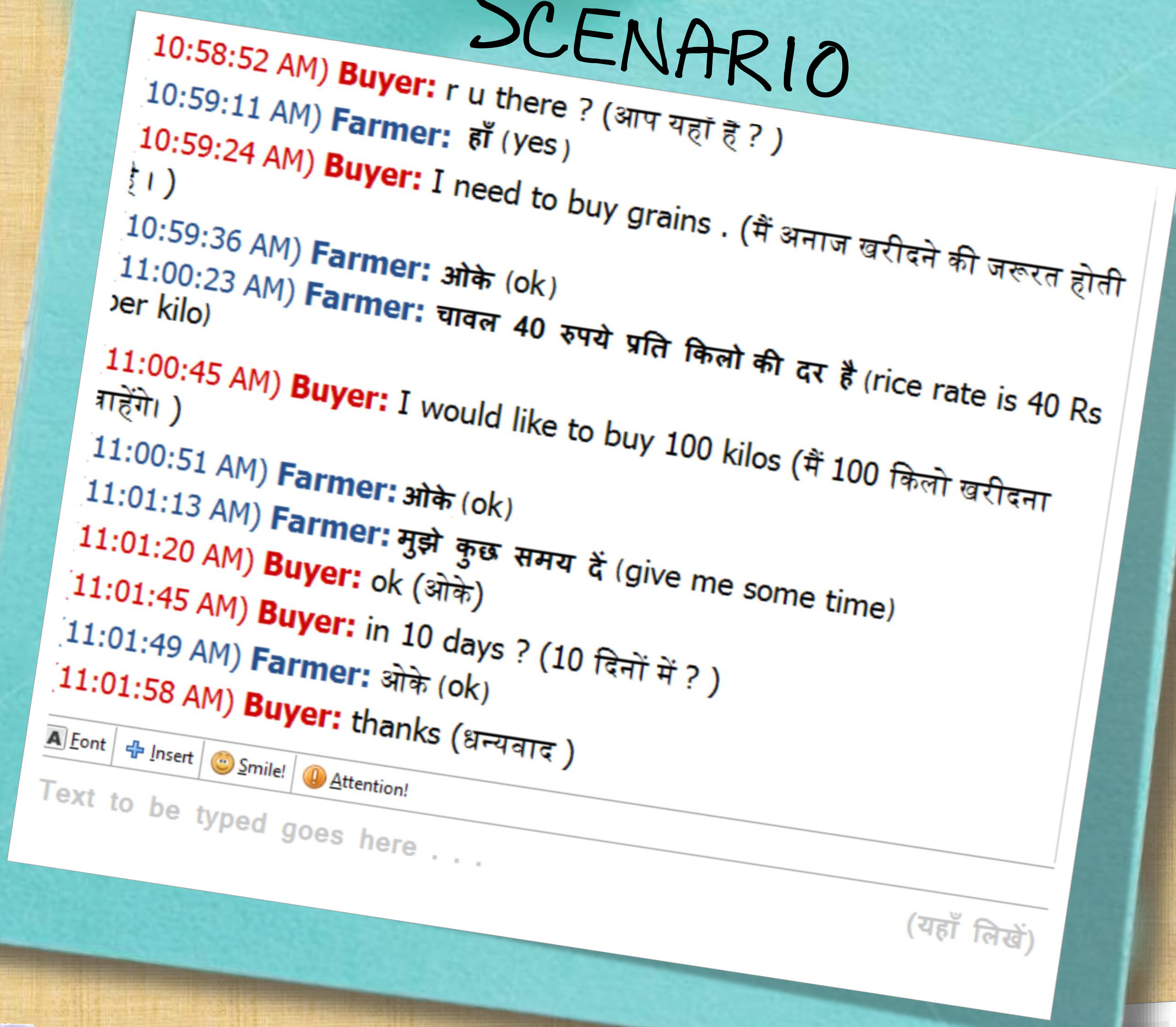
Ovrvw n Motivashn

- TransChat facilitates cross lingual textual communication over English and multiple Indian Languages.
- A client-server IM architecture with multiple Statistical Machine Translation (SMT) engines.

Challngz

- Abbreviated/ungrammatical input has to be converted to linguistically well formed text.
- Language processing modules should be light-weight.

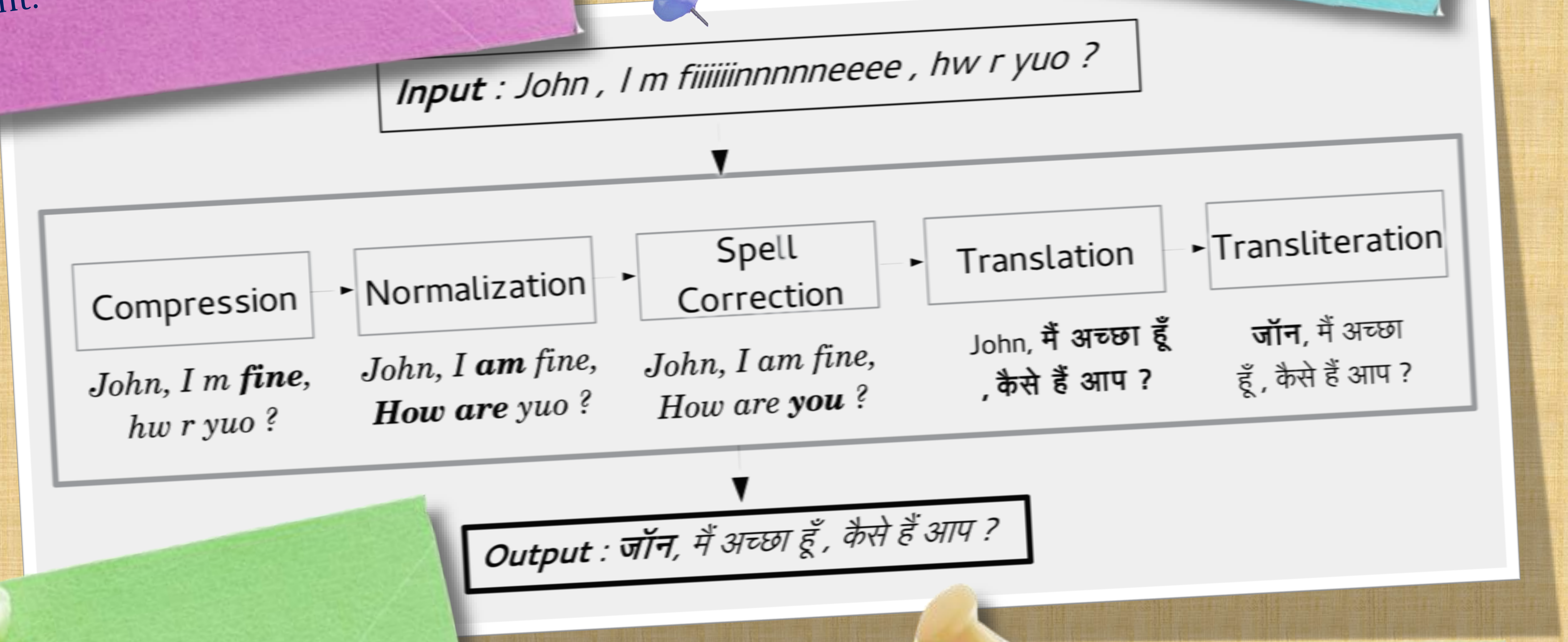
Scenario



10:58:52 AM) **Buyer:** r u there ? (आप यहां है ?)
10:59:11 AM) **Farmer:** हाँ (yes)
10:59:24 AM) **Buyer:** I need to buy grains . (मैं अनाज खरीदने की जरूरत होती
ै।)
10:59:36 AM) **Farmer:** ओके (ok)
11:00:23 AM) **Farmer:** चावल 40 रुपये प्रति किलो की दर है (rice rate is 40 Rs
er kilo)
11:00:45 AM) **Buyer:** I would like to buy 100 kilos (मैं 100 किलो खरीदना
ाहेंगे।)
11:00:51 AM) **Farmer:** ओके (ok)
11:01:13 AM) **Farmer:** मुझे कुछ समय दें (give me some time)
11:01:20 AM) **Buyer:** ok (ओके)
11:01:45 AM) **Buyer:** in 10 days ? (10 दिनों में ?)
11:01:49 AM) **Farmer:** ओके (ok)
11:01:58 AM) **Buyer:** thanks (धन्यवाद)

Text to be typed goes here . . .

(यहाँ लिखें)



Input : John , I m fiiiiinnnnneeee , hw r yuo ?

Compression → Normalization → Spell Correction → Translation → Transliteration

Compression: John, I m **fine**, hw r yuo ?

Normalization: John, I **am** fine, **How are** yuo ?

Spell Correction: John, I am fine, How are **you** ?

Translation: John, मैं अच्छा हूँ , कैसे हैं आप ?

Transliteration: जॉन, मैं अच्छा हूँ , कैसे हैं आप ?

Output : जॉन, मैं अच्छा हूँ , कैसे हैं आप ?

Systm Infrmashn

- Compression: Heuristic based - all the repeated windows of character length greater than two are compressed.
- Normalization: Implemented the normalization system by (Raghunathan and Krawczyk, 2009) as a Phrase Based Statistical Machine Translation system.
- Spell Checking: The JAVA API of Jazzy spell checker is used for handling spelling mistakes.
- Translation: Translations between languages are carried out using Statistical Phrase Based Machine Translation paradigm, powered by the Sata-Anuvadak system.
- Transliteration: We use Google Transliteration API as a post processing step.

Evaluatn Resultzzz

- *Check 1: User experience in terms of (a) using the system and (b) acceptable translation quality etc.*

Three human evaluators asked to give the system 20 messages each. They rate the system by giving (i) Usability Score, (ii) Fluency and Adequacy scores for translated outputs. Scores range from [1-10]. We measure agreement between users through Fliess' Kappa Inter Annotator agreement.

P1-P2	P2-P3	P1-P3
0.63	0.61	0.67

- *Check-2: The pre- and post-processing steps employed helped enhance the quality of translated chat.*

	BLEU (Google)	BLEU (Sata)	BLEU (Bing)	METEOR (Google)	METEOR (Sata)	METEOR (Bing)
Hindi	0.03/**0.19**	0.009/**0.17**	0.041/**0.049**	0.06/**0.23**	0.05/**0.15**	0.07/**0.22**
Marathi	0.004/**0.08**	0.004/**0.01**	N/A	0.03/**0.14**	0.04/**0.08**	N/A
Punjabi	0.01/**0.22**	0.003/**0.04**	N/A	0.04/**0.20**	0.02/**0.09**	N/A
Gujarati	0.04/0.04	0.01/**0.05**	N/A	0.13/0.13	0.06/**0.12**	N/A
Malyalam	0.008/**0.06**	**0.01**/0.008	N/A	0.03/**0.12**	**0.10**/0.06	N/A

Demo

http://www.cfilt.iitb.ac.in/transchat/

Source

https://github.com/cfiltsysads/transchat/

Diptesh कनोजिया[1], Shehzaad ધુલીઆવાલા[1], Abhijit മിശ്ര[1], Naman ਗੁਪਤਾ[2] and Pushpak Bhattacharyya[1]
[1]Center for Indian Language Technology, CSE Department
[1]IIT Bombay, India, and [2]Yahoo Labs, Japan
[1]{diptesh, shehzaadzd, abhijitmishra, pb}@cse.iitb.ac.in
[2]naman.bbps@gmail.com

CFILT
भारतीय भाषा प्रोद्योगिकी केन्द्र