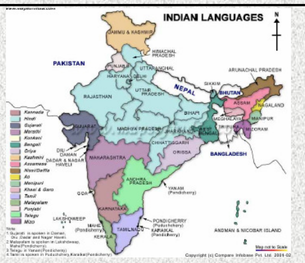
Experiences in Resource Generation for Machine Translation through Crowdsourcing

Anoop Kunchukuttan, Shourya Roy*, Pratik Patel, Kushal Ladha, Somya Gupta, Mitesh M. Khapra, Pushpak Bhattacharyya

IIT Bombay *Xerox Research Centre India



		Speakers
1	Hindi[5]	422,048,642
2	Bengali	83,369,769
3	Telugu	74,002,856
4	Marathi	71,936,894
5	Tamil	60,793,814
6	Urdu	51,536,111
7	Gujarati	46,091,617
8	Kannada	37,924,011
9	Malayalam	33,066,392
10	Oriya	33,017,446
11	Punjabi	29,102,477
12	Assamese	13,168,484
13	Maithili	12,179,122

Context-1 : Linguistic Diversity of India

- India has high degree of linguistic diversity – **22 official languages**, more than 2000 dialects and with large number of users of various languages
- Principal and secondary official languages are **Hindi** and **English**
- Large number of Indians are present on the Internet and in particular, in Crowdsourcing marketplaces
- Several domains have **large translation needs** such as *healthcare, tourism, education, judicial* etc

Session Court
(Regional language)

High Court
(English + Regional language)

Supreme Court
(English)

Context-2 : Judicial Domain in India

- **Multi-tier Judiciary structure** with the Supreme Court, state level High Courts and regional session courts
- The Supreme Court has appellate jurisdiction over High Courts
- Supreme Court proceedings are conducted and recorded in English and in High courts and session courts in respective state languages
- Large translation needs exist to **translate proceedings descriptions from High Courts to the Supreme Court**

Context-3 : Automatic Machine Translation

- Machine Translation is about automatic translation of documents from one language to another
- Statistical Machine Translation (SMT) techniques require **large volumes of parallel corpus** for developing models
- Development of parallel corpora is time consuming, tedious and require participation of expensive linguists

Generate large volume of parallel corpora by Crowdsourcing in a time and cost efficient manner, for developing Statistical Machine Translation systems for judicial domain

Why First-of-a-kind Attempt

- **End-to-end translation system** for judicial domain; in general for domains with non-trivial translational difficulty
- Translation effort in **Indian languages** using Crowdsourcing
- **Large scale** translation effort (of the order of a Million sentences)with no expert involvement
- No use of Gold data (does not exist) and general purpose MT systems (makes life more difficult) to bootstrap

Why Non-Trivial

- Judicial domain **sentences are long, complex and difficult to interpret and translate**
- High degree of **domain specific words** and meanings
- **Sensitive nature of documents** and high potential cost of wrong translation
- **Quality control** in crowdsourced translation is difficult owing to subjective nature of translation work
- High volume of translation requirement as training of SMT systems require of the order of a Million sentences

Sample Sentences

1. *We have heard the learned counsel for the parties in detail and have also perused the writ petition, replies and the precedents relied on by the parties.*
2. *In default of payment of fine, they have been sentenced to undergo rigorous imprisonment for one month each.*
3. *In any case, if the censures were awarded to the petitioner, he should have challenged the same at the appropriate time alleging bias or whatsoever grounds were available to him.*

Sentence Translation Using Social Contacts

Motivation : Explore payment expectation and quality of translation by non experts (having social connections with requestors) for judicial documents

Task : Graduate level course assignment to collect translation for 2K Judicial domain sentences leveraging social contacts; $20 to incentivize crowd

Observation :

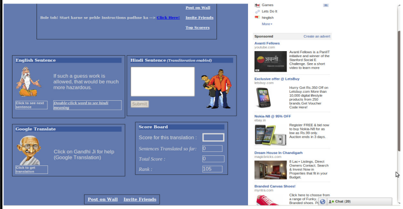
1. Participants found task to be extremely difficult and felt discouraged to participate
2. Interesting approaches such as a Facebook game, *leaderboard* did not help much
3. Participants preferred `push' mode of operation rather than the "pull" mode

Findings:

1. Radical redesign of tasks is needed along with higher level of incentive
2. 10-15 minutes time is required to be spent per sentence
3. The quality of translation is generally acceptable owing to personal connection and responsibility of participants

Data

17K publicly available judicial documents are cleansed and *sentencified* using handcrafted rules to generate 0.5M sentences



Bole toh! Start karne se pehle Instructions padhne ka --> Click Here!

Post on Wall
Invite Friends
Top Scorers

English Sentence

If such a guess work is allowed, that would be much more hazardous.

Click to see next sentence
Double-click word to see hindi meaning

Hindi Sentence *(Transliteration enabled)*

Submit

Google Translate

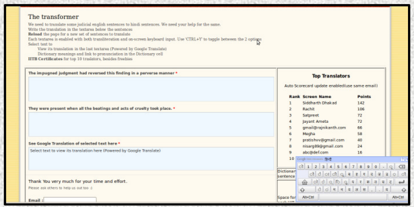
Click on Gandhi Ji for help (Google Translation)

Click to get translation

Score Board

Score for this translation :
Sentences Translated so far: 0
Total Score : 0
Rank : 105

Post on Wall Invite Friends



The transformer

We need to translate some judicial english sentences to hindi sentences. We need your help for the same.
Write the translation in the textarea below the sentences
Reload the page for a new set of sentences to translate
Each textarea is enabled with both transliteration and on-screen keyboard input. Use 'CTRL+Y' to toggle between the 2 options
Select text to
View its translation in the last textarea (Powered by Google Translate)
Dictionary meanings and link to pronunciation in the Dictionary cell
IITB Certificates for top 10 traslators, besides freebies

The impugned judgment had reversed this finding in a perverse manner *

They were present when all the beatings and acts of cruelty took place. *

See Google Translation of selected text here *
Select text to view its translation here (Powered by Google Translate)

Thank You very much for your time and effort.
Please ask others to help us out too :)

Email :

Top Translators

Auto Scorecard update enabled(use same email)

Rank	Screen Name	Points
1	Siddharth Dhakad	142
2	Rachit	106
3	Satpreet	72
4	Jayant Ameta	72
5	gmail@rajnikanth.com	66
6	Megha	58
7	pratishvv@gmail.com	40
8	nisarg89@gmail.com	24
9	abc@def.com	16
10		

Sentence Translation Using Crowdsourcing

Motivation : Explore payment expectation and quality of translation by the *crowd* for judicial documents as well as automatic validation mechanism

Task : 200 Judicial domain sentences with redundancy of 2 for 2 rounds with incentives of $0.1 and $0.2 per sentence in respective rounds. Half of the sentences were to be translated from scratch and remaining ones by modifying machine translation output

Observation :

1. Unacceptably poor quality of translations with 8 % and 32 % of accuracy based on manual validation in respective rounds
2. Higher incentive yielded significant improvement in accuracy but still not good enough
3. Even two correct translations can have significant differences; would be non-trivial task for automatic systems to validate

Findings:

1. Around 10 minutes time had to be spent per sentence and not more than 5 sentences are feasible to translate in one go
2. Mixed response regarding preference modifying MT output over translating from scratch
3. Higher incentives help but Judicial sentences are too long and difficult to attain acceptable level of translation

Questions and Comments for Future

- **Complete automation** (without requiring involvement of linguistic experts) of large scale translation task for specialized domain such as Judicial domain is a reasonably ambitious target to achieve
- **Automatic detection and correction** of poor quality translation is the most difficult hurdle to overcome in crowdsourced translation
  - Identifying spammers is an easy fix but does not help much in resolving the core problem of separating machine translations from human translations
  - Methods using Language Model based features and syntactic parse tree based features are promising but needs further exploration
- **Generation, compared to validation and correction**, is significantly difficult for novice crowd; this is in sharp contrast to preferences of experts
- **User Interface** components are extremely important with emphasis on clear instructions - features like leaderboard, translation aids such as domain dictionary, embedded transliteration etc are helpful

CrowdConf 2011
November 1-2
Mission Bay Conference Center
San Francisco

Presenter & Contact:
Shourya.Roy@Xerox.com