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

**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.

**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) both no expert involvement.
- No use of Gold data (does not exist) and general purpose MT systems (makes life more difficult) to bootstrap.

**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.

**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.

**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.