Experiences in Resource Generation for Machine Translation through Crowdsourcing

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

* Department of Computer Science and Engineering,
IIT Bombay,
{anoopk,kush,somya,miteshk,pb}@cse.iitb.ac.in, pratikpat88@gmail.com

† Xerox India Research Centre,
Shourya.Roy@xerox.com

Abstract

The logistics of collecting resources for Machine Translation (MT) has always been a cause of concern for some of the resource deprived languages of the world. The recent advent of crowdsourcing platforms provides an opportunity to explore the large scale generation of resources for MT. However, before venturing into this mode of resource collection, it is important to understand the various factors such as, task design, crowd motivation, quality control, etc. which can influence the success of such a crowd sourcing venture. In this paper, we present our experiences based on a series of experiments performed. This is an attempt to provide a holistic view of the different facets of *translation crowd sourcing* and identifying key challenges which need to be addressed for building a practical crowdsourcing solution for MT.

Keywords: crowdsourcing, machine translation , parallel corpora

1. Introduction

Wikipedia defines crowdsourcing as the *act of taking a task traditionally performed by an employee or contractor, and outsourcing it to an undefined, generally large group of people, in the form of an open call*[1]. This is a new mode of organizing work which allows large loosely connected individuals to work collaboratively. A crowdsourcing marketplace like Amazon Mechanical Turk (AMT) allows companies or individuals to post jobs for a variety of tasks, attracting millions of users from all over the world. Finally, crowdsourcing cuts the execution cost of tasks by order(s) of magnitude, compared to the traditional means involving experts.

The context for this paper is crowdsourcing and automatic machine translation. Though translation has been known as a specialized domain for linguists, participation of people has always been observed in various forms. The participative culture in translation has appeared in various forms starting from *unsolicited fan translations* [2] to localization of software and digital games by users and gamers respectively to *translation hacking*. O'Hagan (2009) opined that crowdsourcing is the next era of *User Generated Translation* and the more legitimate one. On similar lines, we consider crowdsourcing as a channel for creating linguistic resources in Indian languages towards developing automatic machine translation systems. Machine translation is relatively new in India which started in late 80s and early 90s, with Corpus Based Machine Translation (CBMT) techniques being used quite frequently. The success of CBMT (between a source and a target language) techniques depends on existence of linguistic resources, primarily parallel corpus. A parallel corpus is a large body of source language text and their manual translation in the target language, on which a Statistical Machine Translation (SMT) system is trained. Generating such a corpus is a time consuming and tedious task requiring experts and expensive multi-lingual linguists. There have been sporadic attempts at developing corpora in Indian languages[3] but a radically scalable initiative is required. We question the wisdom of restricting linguistic resource development to a handful of linguists when India has over a billion people and a large fraction of them are familiar with more than one language. Also, some of the languages are spoken by hundreds of Millions of people (approximately, Hindi-Urdu by 500M, Bangla by 250M, Punjabi by over 100M[4]). A prior survey of 733 workers on AMT found that 36% were located in India and most of them are educated with 66% of them having a college degree or higher (Ross et al., 2010). Based on these relevant prior work, we felt that crowdsourcing for linguistic resource development in India has a lot of potential.

Our motivating application comes from the judicial domain in India. India is a heavily multilingual nation with more than 20 recognized languages with Hindi and English as the official and subsidiary official languages respectively. It has a unitary three-tier judiciary, consisting of the Supreme Court, 21 state level High Courts, and a large number of trial courts. Supreme Court proceedings are conducted and recorded in English whereas typically it is done in other rec-

⋆Work done as a Project Engineer at IIT Bombay

[1] http://en.wikipedia.org/wiki/Crowdsourcing

[2] http://en.wikipedia.org/wiki/Fan_translation

[3] http://sanskrit.jnu.ac.in/ilci/index.jsp

[4] http://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers

ognized languages in High Courts, depending on the state. The Supreme Court has original jurisdiction over disputes between the states and the Centre, and appellate jurisdiction over the High Courts. Hence, a large translation need exists, to translate proceedings from High Courts to the Supreme Court, which is presently handled manually. This is partially responsible for justice being delayed - hence denied. The vision is to develop a system towards automatic translation of court proceedings between the Supreme Court and High Courts. This will be a step in providing faster access to information to the courts, thus speeding up the judical process.

This paper captures our experiences from a series of experiments performed on using crowdsourcing for corpus creation. The contribution of this work is:

1. We present a holistic overview of the problem of translation crowdsourcing, identifying the important considerations in the design of a *translation crowdsourcing* system.

2. We report our experiments on crowdsourcing tasks related to English-Hindi machine translation in the judicial domain. Our observations give us hope that crowdsourcing for this challenging translation domain is worthy of deeper exploration.

3. Quality control has already been identified as a major concern, which limits the scalability of crowdsourcing. We decompose the various quality issues, which will help in developing targeted approaches to addressing various quality issues.

4. Given the reach and increasing adoption of social networks in sheer size and diversity, we propose the greater use of social network platforms for crowdsourcing.

2. Related Work

A crowdsourcing marketplace like AMT allows companies or individuals to post jobs for a variety of tasks like review collection (Su et al., 2007), image labeling, user studies, word-sense disambiguation, machine translation evaluation, EDA simulation etc. Snow et al. (2008) explored the use of AMT for generating data for NLP tasks. One of the larger attempts at using crowdsourcing for creating data for human language technologies was the *NAACL-2010 Workshop on Creating Speech and Language Data* with AMT. In this workshop, participants were given $100 to spend on a natural language annotation task of their choice and report their experience. A summary report, Callison-Burch and Dredze (2010) gives snapshots of different natural language tasks taken up by people which ranges from word sense disambiguation to textual entailment to machine translation. Among the machine translation papers from the workshop, an interesting experiential paper (Negri and Mehdad, 2010) presented interesting learning towards reducing verification and translation errors, reducing time etc. especially when there are time and money constraints. Ambati and Vogel (2010) emphasized the importance of quality assurance in crowdsourcing translation. Also, they showed the role of context in phrasal translation which is of significance to us given the long and complex nature of judicial domain sentences.

To reduce the requirement of a large number of manually translated sentences for creation of a parallel corpus, researchers have employed techniques like Active Learning to identify important sentences to translate (Ambati et al., 2010). Such a technique would be helpful to get more quality sentences translated in a shorter time. Callison-Burch (2009) have also explored the use of crowdsourcing for evaluating translation quality of MT systems based on the HTER (Human-mediated translation edit rate) metric. Such an approach provides a good synthesis of human and automated methods, since the utility of automated translation quality metrics is debatable (Callison-Burch et al., 2006). It has been shown that availability of partial alignments for training will makes SMT more effective (Gao et al., 2010). Gao and Vogel (2010) have also demonstrated the use of AMT for collecting alignment from a parallel corpus.

AMT is not the only platform that has been explored for crowdsourcing. von Ahn and Dabbish (2004) demonstrated the use of games to collect data from the crowd. Vickrey et al. (2008) have utilized of games for collecting NLP data. Hu et al. (2010)'s have demonstrated a method to improve a machine translation system by an iterative approach involving only monolingual speakers in the crowd.

3. Experiments in Crowdsourcing

We performed multiple experiments to get insights into various facets of crowdsourcing for developing linguistic resources. In these experiments, we varied different parameters ranging from nature of tasks to type of crowd to incentives offered etc. We describe these experiments and highlight the major observations.

3.1. Low incentive sentence translation (*SentTrans-Course*)

Task: As part of a graduate level course on Natural Language Processing, we floated a course assignment on crowdsourcing for translation. We asked the participating teams to develop English to Hindi translation crowdsourcing applications for the judicial domain. Each team was also allocated a small amount of INR 1000 ($20), which they could use for providing incentives to the crowd. One must note the very low incentive provided as we wanted to know how much we can achieve with such low incentive.

We provided each team with 7000 sentences from the judicial domain, examples of which are shown in Table 1. Though judicial domain sentences are quite long, we selected shorter sentences of an average length of 15 words to make the translation task easier. We asked the teams to experiment around the following themes - crowdsourcing platform, user interaction, spurious translation detection and utilizing the wisdom of the crowds.

Motivation: The motivation for this experiment was to gather experience in building a translation crowdsourcing system with little incentive, while seeking ideas on platform, user interaction and detection of poor quality translations. We also hoped to learn about the demographics of

English	Hindi
Considering the facts and circumstances of the present revision petition, the parties are also left to bear their own costs	वर्तमान पुनरीक्षण याचिका के तथ्यों और परिस्थितियों को देखते हुए, वादियों को स्वयं अपने मुकदमे की फ़ीस का वहन करने के लिये भी छोड दिया गया है
Aggrieved with the aforesaid orders of Financial Commissioner, the present writ petition is filed	वित्तीय आयुक्त के उपर्युक्त आदेश से असंतुष्ट, वर्तमान रिट याचिका की गयी है

Table 1: Judicial domain sentences in English and Hindi

the participating crowd, and the difficulty of translating judicial sentences for non-experts.

Results: From this exercise, we gathered 11361 parallel sentences, which were manually graded into 3 buckets (figures in brackets indicate the number of senteces): good translations (5883), translations containing minor mistakes (1719) and totally wrong or spurious translations (3759). Spurious translation refers to output from another machine translation system, or intentional bad translation and most of such output was completely wrong.

3.2. Sentence translation with AMT (*SentTrans-AMT*)

Task: Next, we floated 200 sentences from the same domain for translation (40 HITs of 5 sentences each) on AMT. Each HIT was floated twice, with a payments of INR 5 ($0.1) and INR 10 ($0.2) per sentence in each round. The incentives were much higher as the sentences to be translated were complex and took significantly longer (around 4 minutes per sentence) than translating general English sentences. The average length of the sentences was 15 words. We did not apply any qualifications for filtering workers.

Motivation: Since AMT is the premier crowdsourcing platform, we wanted to get an idea about generating high quality translations for the judicial domain using AMT. Specifically, we wanted to understand the capabilities of the general crowd for such a task, the quality issues that may be encountered and the incentive structure required. One of the key requirements in translating in a specialized domain like judicial documents is the knowledge of legal terminology and ability to translate legal terminology. We wanted to see if the crowd would be able to translate legal terminology well.

Results: In the first round, at an incentive of INR 5 per sentence, we collected 16 correct, 123 wrong, and 61 spurious translations for the 200 sentences that were floated. At INR 10, we collected 64 correct, 88 wrong, and 52 spurious translations. We observed that there were a small number of contributors who were submitting a lot of spurious translations and exhausting the HITs. We had to float these HITs again after blocking such malicious contributors. Thus, the accuracy of the crowd's translation effort was 8% in the first round, which increased to 32% in the second round. If only the honest responses are considered, the accuracy increases to 11.5% for the first round and 42% in the second round.

3.3. Phrase translation verification with AMT (*PhraseTrans-AMT*)

Task: Based on the experiments, we felt that judicial domain sentences are too long and complex to get a large volume of good quality translations by a non-expert crowd. The final experiment was around phrase translation verification where we extracted 2000 3-word or 4-word phrases from these sentences. We also collected translations for these sentences from a web-based MT system. These phrase translations were floated for *verification* on AMT, as 200 HITs (10 phrase translation pairs per HIT). Each

Category	Observation
Task	Judicial text translation was difficult due to legal terminology. Non-native Hindi speakers would have found Hindi-English translation easier. Time taken per phrase translation verification: 15 s. Time taken per sentence translation (judicial domain): 4 min.
Motivation	Friends and relatives of team members contributed effectively. Many users found translation to be a boring task. The Pareto principle applied applied to workers, about 80% visited the site just once. Teams came up with ideas to encourage translation like hi-scores.
Crowdsourcing Platforms	Near immediate response was seen for tasks on AMT. However, many teams used *Facebook* and custom applications. Accessibility, restricted control over UI were AMT limitations. Many teams chose *Facebook* due to its reach.
Quality Control	Submission of translations from automated MT systems is a problem. A large chunk of spurious translations are submitted by small set of workers. Default choice for translation verification caused spurious submissions. Teams used rule-based methods for detecting spurious translations.
User interaction	Translation aids were provided by teams like transliterated input, Hindi keyboard, bilingual dictionary, legal terminology were helpful to translators.

Table 2: Observations from the crowdsourcing experiments

phrase verification drew an incentive of INR 0.25 i.e. INR 2.5 (5 cents) per HIT. Each phrase was provided an automatic translation and workers had to choose 'Yes' if he agreed with the translation and 'No' if he disagreed. We asked workers to be conservative in agreeing to translations so that only good translations were collected. No sentence context for the phrase was provided.

Motivation: As mentioned, the primary motivation was to see if non experts find verifications easier than sentence translation and produce good quality output. We also observed that, automated web- based MT systems perform reasonably well on translating phrases but not good enough to be considered them directly to train a MT system. In our manual evaluation, 73% (323 out of the 408) of the phrase translations from a web-based MT system were good. This suggested that we could crowdsource verification of these phrase translations and filter out the bad ones.

Results: We did a manual verification of crowd output and found that out of 2000, 301 verifications were correct and these came from 51 valid HITS. We actually collected only 408 verifications, since the remaining 102 verifications were spies inserted to detect spurious submissions. The other 149 HITS were spurious submissions by workers.

Observations from these experiments are listed in Table 2. We elaborate these observations in the following sections, and analyze the various considerations in the design of a translation crowdsourcing system. These considerations are summarized in Table 3.

4. Role of Crowdsourcing Tasks

Machine Translation (MT) systems need different resources depending on the particular MT paradigm (SMT, EBMT, Rule Based MT, etc.) being used. Statistical Machine Translation (SMT) systems need parallel sentence corpora for constructing phrase tables and learning alignments between words, while Example Based Machine Translation (EBMT) requires parallel phrase translations. Hence these are the two tasks we considered for crowdsourcing experiments towards building an MT system for judicial domain. Sentence translation was more difficult than phrase translation, since the former requires source and target language competence, along with an understanding of the domain. The high difficulty level of the task was overwhelmingly raised by teams from the *SentTrans-Course* experiment and respective contributors. This was mostly due to the legal terminology and the complex sentence structures. In addition, participants indicated that they would be more comfortable with Hindi-to-English translation rather than the other way round, and this can probably be ascribed to English being the language of written communication among our target crowd. Secondly, for phrase translation we could design the task as a ***verification*** task and not as a ***creation*** task which also contributed to overall ease of the task. This was clearly borne out of our experiments, where the easier translation verification task yielded better accuracy (78% - 320 out of 408 honest submissions) than the translation task (42% - 64 out of 148 honest submissions) on AMT.

Sentence translation and phrase translation are tasks which naturally arise and can be crowdsourced. Translation is not the only human intelligence intensive task for collecting MT resources. ***Verification*** and ***evaluation of translations*** too requires human intelligence. These tasks are suitable for crowdsourcing. and are important for creating high quality resources and measuring the accuracy of MT systems.

5. Role of Crowd Motivation

We categorize contributor motivations into three broad types, viz. monetary, social and entertainment and analyze how these affect the cost, quality and scale of translation sourcing.

Monetary Gain: In paid crowdsourcing platforms such as AMT, contributors are paid for their efforts. In such crowdsourcing systems which base themselves on monetary payment, the incentive mechanism has to be effectively designed so to be economical and fair. In the *SentTrans-AMT* task, we experimented with 2 different incentive structures for sentence translation, and found that the accuracy of translation increased by 25% as the per sentence incentive increased from INR 5 to INR 10. Owing to the higher complexity and difficulty level of judicial domain sentences, contributors expected higher incentives than typically observed for crowdsourcing tasks in AMT. Also, with increasing incentives people spent more time and effort towards producing better quality translations. However, even at a higher incentive overall accuracy is not satisfactory to use them for training a SMT system. In *PhraseTrans-AMT*, overall cost was lower but usability of collected phrases still needs to be judged.

Social interaction: Today's Web users look for social experiences online in myriad ways and many of them have altruistic goals in contributing to communities in different ways. This is best exemplified by efforts like Wikipedia, open source software development and Facebook internationalization. In our experience, people's social contacts are big contributors to a crowdsourcing effort when there is no incentive or fun element involved. In the *SentTrans-Course* task, we observed that friends, social contacts and relatives actively contributed to the crowdsourcing effort. To attract more contributors, task participants preferred to build their applications on social network portals. Prestige and social visibility are other motivating factors, as the evolution of Wikipedia contributions demonstrates. People are also looking to network with like-minded individuals. Therefore, it is useful to develop a strong community to retain contributors, since there is no obligation on the contributor to keep contributing.

Entertainment: People spend a lot of time online on entertainment activities, and this time can be harnessed to generate useful data. A productive task can be wrapped in the veneer of an entertaining user experience, as demonstrated by the *ESP game* (von Ahn and Dabbish, 2004), which collects image labels using an output-matching two-person game. The user effort comes for free, and hence redundancy can be exploited to ensure quality. The game needs to be interesting enough to retain user loyalty over a long time frame. Game development will take time, however once this fixed cost is incurred the data is available for free. In the *SentTrans-Course* task, we saw some related

Facet	Options	Notes
Task	Sentence/Phrase Translation Translation Verification Translation Ranking Translation Evaluation Alignment Generation	Tasks should be fine grained, easy and uniform. Sentence translation is the most difficult task.
Motivation	Monetary Social Interaction Entertainment Indirect benefit	Uncertainty about data quality is a major spending risk. Game based approaches require detailed design, but data acquisition cost is low. Building a community of contributors is important.
Platforms	Micropayment Marketplace Social Collaboration GWAP Social Networks	Micropayment platforms have a user base and good development tools, but running cost and logistical issues are limitations. Different GWAP paradigms have been developed to ensure high quality data. Innovative ideas are likely to spring from social collaboration. Social networks have a big and diverse reach.
Quality Issues	Junk translations MT system output Bad translations	Submission of other MT systems' ouput is a big problem. Overlap based methods can solve the problem partially. Redundancy required for quality control, but it increases cost.
User Interaction	Transliterated input Bilingual Dictionary Legal Terminology Translation Memory	Translation is a tedious task, hence aids are required. These aids improve productivity and reduce boredom.

Table 3: Considerations for Translation Crowdsourcing

evidences as participants developed features like *Leaderboard* which always shows the top-k leading contributors and thereby motivating people to contribute more to become leaders.

Indirect benefits: People would be willing to contribute effort if they see some indirect benefits. This is seen in collaborative tagging systems like *delicious.com*, where the tags supply data for document clustering while the contributors are benefited by recommendations. Generally, attention spans of contributors would be short in these cases. Such users would be willing to contribute to tasks which do not involve too much effort. Quality of the data generated cannot be guaranteed, but aggregating and averaging data over a large number of users will help improve the data quality. This seems to be the approach behind the *Duolingo* [5] project, where people contribute translations during the process of learning a language.

6. Role of Crowdsourcing platforms

While crowdsourcing is an evolving technology, a few paradigms and platforms for development of crowdsourcing systems are crystallizing. We survey the major crowdsourcing platforms.

6.1. Microtask Platforms

Most of the current research (Callison-Burch and Zaidan, 2011; Munro, 2010) on leveraging crowdsourcing for NLP and translation has used microtasking platforms like AMT, CrowdFlower, MicroWorker, etc. In AMT, each *requester* posts *Human Intelligence Tasks* (HITs), inviting *workers* to work on those. Workers submit the completed HITs, and after the requester's approval the worker receives a prespecified *reward*. Similar practices are followed in other popular platforms.

The appeal of these microtasking platform lies in the access to their large user base, the ease of development and deployment of tasks, and the ease of payment and worker management. Consequently, the setup time is reduced, and the dynamics of the marketplace would hopefully drive the costs down. The platforms can be said to operate in *pull mode*, where workers discover tasks reducing the promotion burden on the requester. We got a near instantaneous response from workers on the *SentTrans-AMT* and *PhraseTrans-AMT* experiments. The microtask platform vendors take a cut from the payments as the commission for these facilities.

For the *SentTrans-AMT* experiment, we got good translations for 42% of the honest submissions at an incentive of INR 10. For the proper incentive, workers are willing to put in effort which is encouraging for a difficult domain like judicial translation. To make the system economically efficient, it is necessary to choose the most important sentences for translation as proposed by Ambati et al. (2010).

However, these platforms have some limitations which were observed by the teams in the *SentTrans-Course* experiment. Crowdsourcing platforms have limited payment options in terms of currency and modes of redemption. Control over user interface and job distribution is limited. Geographical restrictions on workers, and legal hassles are an-

[5] http://duolingo.com

other limitations. Hence, 8 out of 11 teams preferred to build custom applications. On the flip side, a custom platform will operate in *push mode*, where the developers will have to take the responsibility of attracting contributors and promoting the platform, as participating teams discovered.

6.2. Games with a Purpose Platform

The use of games for crowdsourcing for collecting annotated data has been pioneered by von Ahn and Dabbish (2004). Subsequently, von Ahn (2006) has shown the effectiveness of games in a variety of annotation tasks. These games are multi-player, anonymous games designed around paradigms of output-matching, role-inversion, input-matching in order to ensure high quality labels and prevent collusion (von Ahn and Dabbish, 2008). Though the principles for design of such games are evolving, there is no generic framework available for building GWAP applications. A very good game design and interaction model needs to be put in place to retain loyalty and interest. A game design involves substantial effort and it may be worthwhile only if a large volume of data needs to be generated. Incorporating social aspect into games like having multi-person competition games, showing leader-board scores generates more interest.

There have been a few efforts to develop games for collecting NLP annotations. Vickrey et al. (2008)'s games for collecting categorical relations and selectional preferences of verbs is an example. *Verbosity* and *Pletch* are good examples of games for collecting NLP annotations. We know of no work on developing games for translation crowdsourcing. Designing a game for the *translation* task may be difficult, but simpler tasks like *verification* and *alignment generation* look like good candidates for developing crowd-pulling games. Recent work from Hacker and von Ahn (2009) proposes the use of games for collecting preferences, and this approach may be applicable for the *translation ranking* task.

6.3. Social Collaboration Platform

There are collaborative efforts for generating community content, often with altruistic intentions. Tasks may be simple like looking for objects in images to locate missing people, or complex ones like building software (Open Source Software) or building an encyclopedia (Wikipedia). As the task complexity increases, the effort to maintain the system increases requiring higher levels of co-ordinations among the members. Generally, a small set of active, highly motivated contributors take up this responsibility and drive the community. Such communities take time and effort to build, but often are the sources of good ideas and breakthroughs. As an example, in a novel exercise, Munro (2010) used crowdsourced translations of SMSes to aid rescue efforts in the aftermath of the Haitian earthquake of 2010.

6.4. Social Networks

Social networking has seen an explosive growth over the last few years and this has resulted in a platform that has a large and diverse reach. It can complement the other crowdsourcing platforms by providing access to large crowds with varied linguistic capabilities. For instance, microtasking over a social networking platform looks to be an interesting prospect. However, *Facebook*, the largest social network, has not evolved into a viable enough system for microtasking yet. Games are wildly popular on Facebook, thus providing an ideal platform for GWAP based translation crowdsourcing. Causes go viral on social networks, so appeals to altruism can be spread fast on social networks.

7. Role of Quality Control

Obtaining translations of high quality is one of the primary concerns in an translation crowdsourcing system. A couple of questions related to quality need to be pondered over first. First, is it possible to create expert quality translations using a crowd of non-experts? 42% of the honest translations we collected from AMT for the judicial domain were good, which is encouraging considering the complexity of the domain and the absence of any redundancy. Two, is verification of generated translations an easier task than automated translations? Automated evaluation of MT output has so far proved to be a tough problem, with opinion still divided over MT evaluation techniques (Callison-Burch et al., 2006). Hence, it would be overly optimistic to expect highly automated verification of translations.

However, there may be some classes of translation quality issues that are easy to detect. We have identified the following quality issues, with the hope that these can be individually targeted for automatic verification:

- Wrong punctuations, presence of decimal numbers from the English Unicode character set, etc., which are minor issues that can be ignored.

- Junk translations, which are translations submitted by the crowd which have no relation to the source sentence. Such submissions may not have any overlap with the source sentence or be just a repetition of few words. Teams used bag-of-words based overlap method or a few rules to identify these sentences. These methods obtained good precision.

- Contributors resort to submitting translations from Web-based MT systems frequently, which are likely to be inaccurate. This category contributed a substantial 25% of the sentence translations we collected in the *SentTrans-AMT* experiment. A majority of the errors in these spurious translations are related to tough NLP problems like agreement, matching case-markers, multi word translation, and word sense disambiguation. Hence, it is difficult to automatically detect such spurious translations. Contributors may also do some post editing of these translations to avoid detection, hence a simple string match against known Web based MT systems will not be able to find these spurious translations. In the *SentTrans-Course* experiment, some teams used methods based on overlap and edit distance of submitted translations with automated translations, and these methods achieved high precision but were low on recall.

- The translation provided by the user may not be good. Since, the contributors are not expert translators, we expect a lot of such translations to be submitted. The

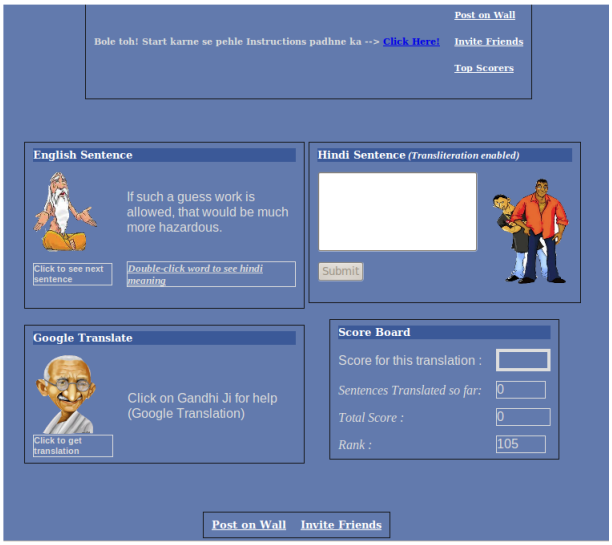

Figure 1: A Translation Crowsourcing application on *FaceBook*

main problems we found with the translations collected from our experiments were mistakes in choice of word senses, and use of case markers. It is the primary challenge of MT crowd sourcing to identify good translations and build a near expert quality parallel corpus in such an environment. Callison-Burch and Zaidan (2011) propose a system based on collecting translations, their edits, and ranking judgments on translations and training a model to score the translations from extracted features. At the heart of this method is the use of redundancy, and we believe that any method to identify good translations will have to use redundancy, since the task is subjective. However, redundancy increases the translation cost.

8. Role of User interaction

All teams observed that the *Pareto* principle applied to the people visiting their application, with only about 20% visiting the application again. Providing the right user interaction is an important step to convert a visitor to a regular contributor. Teams incorporated a number of features in their translation systems to provide a good user experience. Figure 1 shows the user interface developed by one particpating team. Typing in a non-Roman script is difficult for most people, as they wouldn't be familiar with alternative keyboard layouts. Teams provided a transliteration system for input, which greatly encouraged contributors. Translation being a tedious task, teams also provided translation aids like dictionary lookups and translation memories. We also observed that people were more comfortable providing English translations to Hindi sentences rather than the other way round. This is presumably because English is the dominant language of written communication. This indicates that, if possible, the target language should be the dominant language of written communication. Translation guidelines help the contributors make the right choices and assumptions in case of doubts. Modeling the tasks as games, and incorporating competitive elements like leader-boards and time-limited games can increases user interest. User interaction also has an help on discouraging spurious submissions. In the *PhraseTrans-AMT*, the verification task had a default choice, which encouraged spurious submissions.

9. Conclusion

Our experience has helped us get a holistic view of crowdsourcing, a summary of the various facets of crowdsourcing is depicted in Table 3 We summarize our learnings in the form of a SWOT analysis of crowdsourcing for translation.

- **Strengths:** Crowdsourcing offers access to a large, educated user base which can be used to create corpora on a scale not previously known for Indian language machine translation. Our experience in judicial domain crowdsourcing is encouraging. As Web 2.0 technologies have shown, the wisdom of the crowds can be relied upon to generate expert quality content. Effects of scale make crowdsourcing cost-effective.
- **Weaknesses:** Ensuring quality of translations is the biggest concern, since it is a collection of non-experts who are generating data. Figuring out the right model to generate expert quality data is necessary. More important is the issue of spurious translations submitted by the crowd, and effective mechanisms are needed to detect them. We have identified various kinds of quality issues, which can be focused on.
- **Opportunities:** Most of the work on crowdsourcing translations has relied on crowd marketplaces, but we should also actively explore other modes like games with a purpose and social collaborative work projects and games for generating translations. Social network as a platform for crowdsourcing is an encouraging prospect.
- **Threats:** Will crowdsourcing marketplaces be cost-effective in the long run? Today supply of workers far exceeds demand, but that may change over time.

10. Acknowledgements

We would like to thank Nicola Cancedda for his valuable inputs during the discussions.

11. References

Vamshi Ambati and Stephan Vogel. 2010. Can crowds build parallel corpora for machine translation systems? In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

Vamshi Ambati, Stephan Vogel, and Jaime Carbonell. 2010. Active learning and crowd-sourcing for machine translation. *Language Resources and Evaluation LREC*, 11.

Chris Callison-Burch and Mark Dredze. 2010. Creating speech and language data with amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

Chris Callison-Burch and Omar Zaidan. 2011. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the NAACL HLT 2010*

Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *Proceedings of EACL-2006*.

Chris Callison-Burch. 2009. Fast, cheap, and creative: evaluating translation quality using amazon's mechanical turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*.

Qin Gao and Stephan Vogel. 2010. Consensus versus expertise: a case study of word alignment with mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

Qin Gao, Nguyen Bach, and Stephan Vogel. 2010. A semi-supervised word alignment algorithm with partial manual alignments. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*.

Severin Hacker and Luis von Ahn. 2009. Matchin: eliciting user preferences with an online game. In *Proceedings of the 27th international conference on Human factors in computing systems*.

Chang Hu, Benjamin B. Bederson, and Philip Resnik. 2010. Translation by iterative collaboration between monolingual users. In *Proceedings of Graphics Interface 2010*.

Robert Munro. 2010. Crowdsourced translation for emergency response in haiti: the global collaboration of local knowledge. In *AMTA Workshop on Collaborative Crowdsourcing for Translation*.

Matteo Negri and Yashar Mehdad. 2010. Creating a bilingual entailment corpus through translations with mechanical turk: $100 for a 10-day rush. In *Proceedings of the NAACL-2010 workshop on "Creating Speech and Language Data With Amazons Mechanical Turk"*.

Minako O'Hagan. 2009. Evolution of user-generated translation: Fansubs, translation hacking and crowdsourcing. *The Journal of Internationalisation and Localisation*.

Joel Ross, Lilly Irani, M. Six Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who are the crowdworkers?: shifting demographics in mechanical turk. In *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems*.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Qi Su, Dmitry Pavlov, Jyh-Herng Chow, and Wendell C. Baker. 2007. Internet-scale collection of human-reviewed data. In *Proceedings of the 16th international conference on World Wide Web*.

David Vickrey, Aaron Bronzan, William Choi, Aman Kumar, Jason Turner-Maier, Arthur Wang, and Daphne Koller. 2008. Online word games for semantic data collection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Luis von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*.

Luis von Ahn and Laura Dabbish. 2008. Designing games with a purpose. *Commun. ACM*, 51, August.

Luis von Ahn. 2006. Games with a purpose. *Computer*, 39(6), june.