Human Computation and Crowdsourcing: Works in Progress and Demonstration Abstracts
AAAI Technical Report CR-13-01

English to Hindi Translation Protocols for an Enterprise Crowd

Srinivasan Iyengar, Shirish Karande, Sachin Lodha
TCS Innovation Labs, 54-B Hadapsar Industrial Estate, Pune 411013, India
shirish.karande@tcs.com

Abstract

We present early results on crowd sourcing translations in an enterprise setting. We show that several weak translators can together converge to translation quality higher than individually plausible. We share learning about post editing of translations and the effort perceived by the crowd. A key observation is that a protocol "*machine-human-human*" that utilizes two-hop post editing can provide an almost expert quality translation. Finally, we discuss methods that can choose the best translation among several candidates.

Introduction

Current quality of machine translation for Indian languages is not satisfactory. However, statistics point towards a promise of crowd sourcing for Indian Languages: (Ross et. al., 2010) have observed that as of February 2010, 46% of MTurkers on Amazon are from India. Moreover, Indian IT enterprises employ large number of educated computer savvy Indians (see Table 1) and are well placed to utilize these employees for crowd sourcing activities. Motivated by these observations, we present early results of soliciting English to Hindi translations, under various collaborative mechanisms, from a crowd within an Indian IT enterprise.

Table 1 Employees in Indian IT Companies (source Wikipedia)

Enterprise Name	No. of Employees
Tata Consultancy Services	276,196
Cognizant Technology Solutions	162,700
Infosys	155,629
WIPRO Ltd.	140,569

Related Work

Post-editing and redundancy can improve translation quality; however, at the cost of efficiency. Consequently, coordination of collaborative translation is increasingly being investigated. The Monotrans project (Hu et. al., 2010) has investigated protocols to involve monolingual speakers to produce translations [5]. (Ambati et. al., 2012) break crowd sourcing into three phases: the translation is first given to a weak bilingual, then edited by bilinguals and finally by monolinguals. Other examples of recent work on post-editing are: (Liao et. al, 2011, Green et. al., 2013). Our work builds upon these efforts, specifically for an enterprise setting. It is reasonable to assume that in an enterprise, several checks to discourage spammers are in place. Such assumptions can alter the crowd coordination as well as the strategy for identifying the best translation.

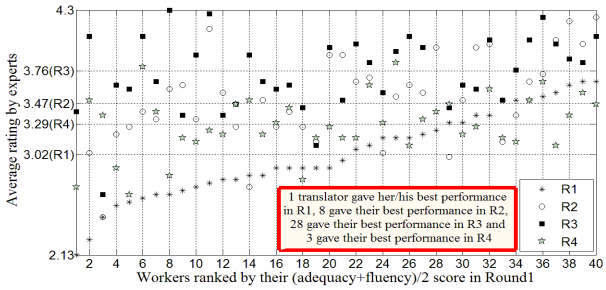

Figure 1 The average rating obtained by the translator in each round

Data Collection

We used 40 employees to translate 200 sentences sampled from the 2005 ACL shared task. The sentences were sampled based on the word length distribution among the 100 most popular articles on Wikipedia (in 2012). The translations were solicited over 4 rounds, such that each worker contributed to 15 sentences per round, thus contributing to a total of 2400 translations. The sentence allocation was semi-random to avoid significant cliques:

- **R1:** Translate the English sentence without the help of any machine or human translation as reference
- **R2:** Edit the translation from Google Translate.
- **R3:** Edit the translations obtained in the second round. Each translation was edited by only one worker.
- **R4:** Choose and edit the 3 Hindi translations obtained in R1. The English sentence is not disclosed.

In each round, we asked the crowd to indicate the effort perceived by providing a score in the range 1 to 10. In addition, we got expert rating on the following parameters:

- **Adequacy:** (5) all (4) most (3) much (2) little (1) no meaning
- **Fluency:** (5) flawless (4) good(3) non-native (2) disfluent (1) incomprehensible

Quality of the Enterprise Crowd

The worker performance can be observed in Figure 1. In R1 the adequacy (fluency) was 3.05 (2.99). In comparison the ratings for Google Translate was 1.82 (1.78). Thus,

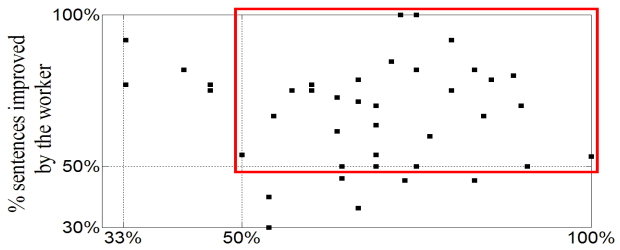

Figure 2 Performance in R2 as a translation provider v/s as an editor.

even without any machine assistance, the crowd is likely to provide a translation which is non-native yet fluent and captures most of the meaning, whereas a machine translator would be disfluent with little meaning. R2 showed an average improvement of 15% over R1. The crowd's ability to improve the machine translation varied with the pre-edit quality. However, on average all workers improved the translations with pre-edit rating less than 3. R3 exhibited a gain of 8% over R2. The performance in R3 shows that even weak translators can often improve the translation by a better translator. Figure 2 shows that atleast 33% of each worker's sentences were improved upon, and every worker improved atleast 30% translations. Moreover, nearly 75% of the crowd belonged to the quadrant that represents mutual improvement. This implies that translators needn't be asked to edit sentences in an order strictly influenced by their quality. R4, which emulates the performance of monolinguals, exhibited a 9% improvement over R1. Finally, we observed that the average effort perceived by the crowd was 5.5, 4.93, 2.9 and 2.52 in rounds 1 to 4. Thus, side-information not only improves the quality but can reduce the effort by 54%.

Quality Translation using Edit Distances

In order to differentiate among the translations, we employ a sentence similarity measure (described below) that builds upon a modified Damerau-Levenshtein distance for word comparison. The modified distance treats phonetically similar characters as same. We consider several strategies for translation selection based on this measure. The efficacy of our strategies is measured by comparing the crowd's mean performance in a round with the quality of our choices for a round averaged over all sentences.

Let, S_j^x represent the j^{th} translation in round x, with S^0 being machine translation. In addition, let the j^{th} translation in R3 be obtained by editing the j^{th} translation in R2. Then our selection criterion for each round can be described as:

$\boldsymbol{Similarity(\alpha)}$: $\boldsymbol{Input}$: *Sentences* S_1, S_2 *s.t.* $|S_2| \geq |S_1|$

$\boldsymbol{Let, for\ every}$ $s_{1,j} \in S_1,\ s_{2,k} \in S_2$

$$\omega_{j,k} = min\{|s_{1,j}|, |s_{2,k}|\} / max\left\{\sum_{l=1}^{|S_1|}|s_{1,j}|, \sum_{p=1}^{|S_2|}|s_{2,k}|\right\}$$

$$W_1 = GetSynsets(s_{1,j}),\ W_2 = GetSynsets(s_{2,k})$$

$$m_{j,k} = 1 - \min_{w_a \in W_1, w_b \in W_2}\left\{\frac{Modified\_DL(w_a, w_b)}{max\{|w_a|, |w_b|\}}\right\}$$

$$\sigma = MaximalBipartiteMatching([m_{j,k}])$$

$$\alpha = \sum_{l=1}^{|S_1|}\left((1 + \delta[\sigma(l) - \sigma(l-1) - 1])m_{l,\sigma(l)}\omega_{l,\sigma(l)}\right)/2$$

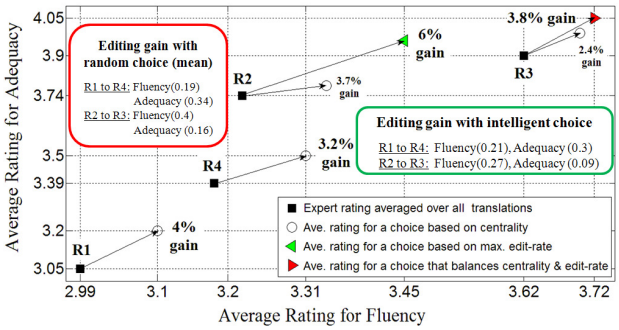

Figure 3 Ratings for selected translations, averaged over all sentences

R1: $S^{1*} = \arg\max_{S_j^1}\left\{\sum_{k=1:3}\alpha(S_j^1, S_k^1)\right\}$ **R2:** $S^{2*} = \arg\min_{S_j^2}\left\{\alpha(S^0, S_j^2)\right\}$

R3: $S^{3*} = \arg\max_{S_j^3}\left\{\left(1 - \alpha(S^0, S_j^2)\right)\cdot\alpha(S_j^2, S_j^3)\right\}$

R4: $S^{4*} = \arg\max_{S_j^4}\left\{\sum_{k=1:3}\left(\alpha(S_j^4, S_k^1) + \alpha(S_j^4, S_k^4)\right)\right\}$

The strategy in R1 is motivated by consensus (centrality). In R2 we chose the candidate with the highest edit-rate (novelty). Edit-rate alone works iff the worker quality is high, and the pre-edit quality is low. Hence, in R3, we seek to balance novelty with centrality; meanwhile, in R4, as all workers get the same side-information, the R1 translations are also included in evaluation. Figure 3 shows that our strategies provide 3-6% gain over the mean performance.

Conclusion

The quality of an enterprise crowd is significantly better than machine translators. Weak translators and monolinguals often improve the quality of better translators. Moreover the crowd can achieve expert quality, with a mechanism that uses machine translation as input with multiple hops of human edits, along with a translation selection strategy that balances consensus with novelty.

References

Ross, J., Irani, L., Silberman, M., Zaldivar, A., & Tomlinson, B. (2010, April). *Who are the crowd workers?: shifting demographics in mechanical turk*. In Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems (pp. 2863-2872). ACM.

Hu, C., Bederson B. B, and Resnick, P., *Translation by iterative collaboration between monolingual users*. In Proceedings of the ACM SIGKDD Workshop on Human Computation, HCOMP '10, pages 54–55. ACM, 2010.

Ambati, V., Vogel, S., & Carbonell, J. (2012, February). *Collaborative workflow for crowd sourcing translation*. In Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work (pp. 1191-1194). ACM.

Liao, S., Wu, C., & Huerta, J. (2011). *Evaluating human correction quality for machine translation from crowdsourcing*. Recent Advances in Natural Language Processing (RANLP 2011), 598-603.

Green, S., Heer, J., & Manning, C. D. (2013). *The Efficacy of Human Post-Editing for Language Translation. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 439-448). ACM.*