On Transcribing Russian with a Highly Mismatched Indian Crowd

Purushotam Radadia, Shirish Karande, Sachin Lodha
TCS Innovation Labs-TRDDC, Pune, India
purushotam.radadia@tcs.com, shirish.karande@tcs.com, sachin.lodha@tcs.com

Abstract

We present early results on the possibility of transcribing non-Indian languages using a rural Indian crowd that types in vernacular scripts. Here, we present results for transcribing Russian in Gujarati, Marathi and Telugu scripts without having any understanding of Russian. We observe that this highly mismatched crowd is able to achieve non-trivial accuracy. We show that one can effectively combine crowd work across four non-native languages to get a word recognition rate of ~55% and 4-best list recognition of ~71%

Introduction

The increased mobile penetration in rural India, can be used to derive a demographic advantage for crowd work. It is estimated that there are 75 million internet users outside the top 30 cities. Given the popularity of several OSNs a Voice reCAPTCHA service attached to any one of them could potentially capture millions of transcriptions in a short time span. However, rural Indians are typically educated in vernacular languages and often read/write only in their mother tongue. Thus, such a highly mismatched crowd may appear of little value for transcribing non-Indic languages. In this work we set about to dispute this notion. We observe ~25% accuracy averaged over 141 users and three Indic languages. We show that publicly available transliteration engines can be used to combine/compare crowd work across multiple Indic language scripts to achieve an average accuracy of 42%.

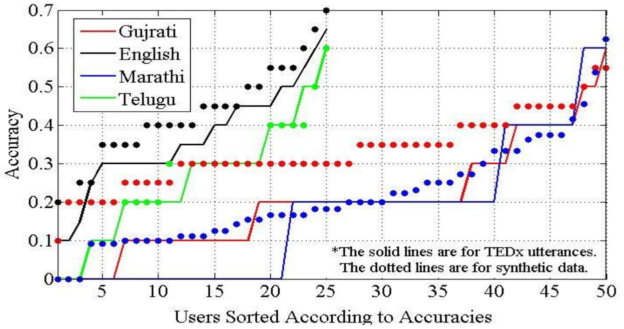

Figure 1 Variance in user accuracy despite not knowing Russian

Related Work

Several studies (e.g.Audhkhasi et. al. 2011) have shown the utility of transcriptions from a non-expert crowd. However, recent work by (Jyothi et. al. 2015) is the first to investigate the use of a mismatched crowd, i.e., a worker that is unfamiliar with the source language. Nevertheless, in their study a user types the response in English, we consider a high degree of mismatch by collecting response in Indic language scripts. Further, we use Russian as the source language.

Data Collection

Our corpus consists of two sets (I) Natural (N) that contains 250 words isolated from Russian TED talks (II) Synthetic (S) that contain 500 phonetically rich words synthesized using Google. The (S) set consists of 150 short (2-5 arpabets), 200 medium (8-10 arpabets) and 150 long (14-16 arpabets) words sampled from a large Russian Pronunciation dictionary[1]. Similarly it was observed that (N) set consists of 84 short words (N-S) and 166 words (N-M) with 6-13 arpabets.

We utilize the following crowd for tasks of 15-40 words:
Gujarati: 66 students (8th-10th) from Pithadiya, Saurashtra.
Marathi: 50 students (7th-9th) from municipal school, Pune.
Telugu: 25 villagers from Alavalapadu, Kadapa district.
English: 31 volunteers from an IT company.

The user listened to a Russian word and typed the response in his/her configured "*native*" script. Table 1 shows sample transcriptions of word *клюква*. We collect 6 transcriptions for each word (2 Gujarati, 1 Marathi, 1 Telugu, 2 English).

Table 1: User responses of word *клюква*

English	Gujarati	Marathi	Telugu
kluckwah	કલુકવા	कॉकवाय	కులుకువ

Table 2: Average crowd performance

	N	N-S	N-M	S	S-S	S-M	S-L
Gujarati	0.20	0.24	0.21	0.33	0.25	0.37	0.35
Marathi	0.17	0.13	0.18	0.21	0.15	0.26	0.27
Telugu	0.26	0.16	0.29	0.29	0.25	0.24	0.40
English	0.37	0.27	0.42	0.43	0.32	0.49	0.51

[1] Dictionary: http://sourceforge.net/projects/cmusphinx/files/Acoustic and Language Models/Russian/

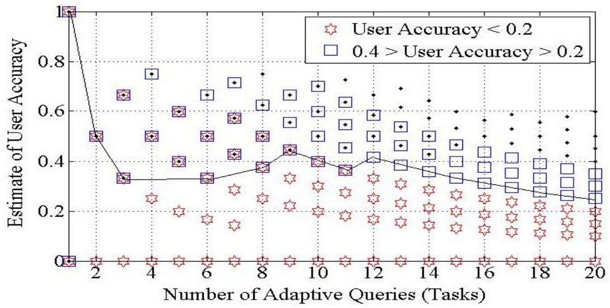

Figure 2 Adaptive task allocation for early rejection of workers

Experimental Methodology

We use English as a common script to compare the user responses and Russian ground-truth. We transliterate Russian with *read phonetically*[2] and Devnagari responses with-*Pramukh*[3]. We evaluate the decoding accuracy against a limited dictionary of 5000 words created by randomly sampling the Russian dictionary. To decode a transcription we evaluate a word length normalized Levenshtein distance, this was observed to provide performance comparable to ROVER (Fiscus, 1997). Finally, for improved decoding, we used the Carmel Finite State Transducer toolkit[4] to map the Russian arpabets to the transliterated English characters.

Crowd Performance

Table 2 shows that the accuracy improved as the length increased. The average accuracy provided by the Gujarati, Marathi and Telugu crowd was ~25%, which is lesser than the mismatched English crowd. This difference in performance can be attributed to the fact that the crowd from an IT company is likely to be already pre-filtered along possibly favorable parameters (e.g. Iyengar et. al. 2013). In future, it may be feasible to apply pre-filtering for all workers. For example, historical accuracy associated with transcribing a word can be used to design an adaptive word test. Figure 2 shows the result of simulated adaptive tests on 50 Gujarati users. We can identify users with less than 10% (20%) accuracy in at most 8 (12) adaptive transcription requests.

In addition to pre-filtering (*channel selection*) there can be many other ways to improve mismatched crowd accuracy:

<u>Combining Work</u> (*parallel-channels*): Table 3 shows the performance improvement obtained by combining multiple transliterations. Combining two Gujarati transliterations improves performance by ~5% over a single Gujarati transcription. The results of combing 2 English transcribers are more impressive, however, we can observe that a similar effect can be obtained by combining more number of Indic transliteration. Furthermore, the English combination can be improved by including inputs from the Indic transliteration.

<u>Phoneme Mapping</u> (*channel modeling*): The improved phonetic mappings leads to a performance gain of (3-8%) for all the Indic language combinations. Additional data and better noise handling during training could improve these results. The use of FSTs seems to actually worsen the performance for English. This can be attributed to the higher accuracy of Google Translate in transliterating Russian to English v/s our mapping from English to Arpabet.

Table 3: Combining crowd work

	N	N-S	N-M	S	S-S	S-M	S-L
2-Guj	0.24	0.23	0.24	0.39	0.33	0.42	0.41
2-Eng	0.42	0.29	0.48	0.55	0.39	0.63	0.63
4-Indic	0.35	0.29	0.38	0.48	0.37	0.53	0.55
6-All	0.47	0.42	0.49	0.58	0.47	0.62	0.65

Table 4: Impact of phoneme mapping using FST

	Guj	Mar	Tel	Eng	2-Guj	2-Eng	4-Ind	6-All
N	0.25	0.2	0.29	0.33	0.3	0.46	0.42	0.5
S	0.36	0.24	0.31	0.41	0.47	0.52	0.54	0.6

Table 5: Performance in terms of 4-best lists

	Guj	Mar	Tel	Eng	2-Guj	2-Eng	4-Ind	6-All
N	0.33	0.28	0.36	0.49	0.49	0.58	0.50	0.62
S	0.43	0.29	0.40	0.58	0.52	0.76	0.61	0.74
N	0.45	0.38	0.52	0.58	0.5	0.64	0.57	0.66
S	0.55	0.41	0.55	0.60	0.6	0.7	0.66	0.75

*The shaded rows show the performance without training a FST.

<u>*4-Best Lists*</u> (*soft decoding*): Some transcription protocols may require *n* likely word transcriptions rather than an exact decoding. Table 3 shows that such a relaxation can lead to a typical performance gain of more than 10%. Furthermore, one can now note that Marathi and Telugu crowd have benefited a lot more with the use of FSTs for phonetic mapping.

Conclusion

We observe that Vernacular rural Indian crowd can effectively help in transcription protocols and transcriptions in different scripts can be combined to improve accuracy.

References

Audhkhasi, K., Georgiou, P., & Narayanan, S. S. (2011, May). *Accurate transcription of broadcast news speech using multiple noisy transcribers and unsupervised reliability metrics*. In Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on (pp. 4980-4983). IEEE.

Jyothi, P., & Hasegawa-Johnson, M. (2015, February). *Acquiring Speech Transcriptions Using Mismatched Crowdsourcing*. In Twenty-Ninth AAAI Conference on Artificial Intelligence.

Fiscus, J. G. (1997, December). *A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER).* In Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on (pp. 347-354). IEEE.

Iyengar, S., Karande, S. S., & Lodha, S. (2013, March). *English to Hindi Translation Protocols for an Enterprise Crowd*. In First AAAI Conference on Human Computation and Crowdsourcing.

[2] Read Phonetically, https://translate.google.co.in/

[3] Indic Language Typing: http://service.vishalon.net/pramukhtypepad.aspx

[4] J. Graehl: http://www.isi.edu/licensed-sw/carmel/