Quality assessment of crowdsourcing transcriptions for African languages

Hadrien Gelas[1,2], Solomon Teferra Abate[2], Laurent Besacier[2], François Pellegrino[1]

[1]Laboratoire Dynamique Du Langage, CNRS - Université de Lyon, France, [2]Laboratoire Informatique de Grenoble, CNRS - Université Joseph Fourier Grenoble, France

The Big Question

- Amazon's Mechanical Turk (Mturk) is proved to be powerful for many NLP tasks
- But is it true for all languages ?
- Our goal is to evaluate the quality of crowdsourcing transcriptions for African Languages
- No massive data collection (we already had the transcriptions we wanted to transcribe)
- Ethical issues also discussed

Starting point

Speech transcription

- Essential in speech recognition (ASR) systems
- Tedious and expensive

Mturk

- Online market place for work
- Mass of workers always available: fast accomplishment and low payment rate
- Great potential to speed up and reduce the cost of many NLP tasks !

Related Work

- Many recent studies investigated the use of Mturk for various NLP tasks, among them, speech transcriptions: [1, 2, 3, 4, 5]
- Got fast and cheap near-expert accuracy transcriptions
- But, most of them concentrate on English
- No investigation for African languages !

Languages

Amharic

- Ethio-Semitic language
- Over 22 millions speakers (17m natives)
- Own syllabary writing system

Swahili

- Bantu language
- Over 50 million speakers (5m natives)
- Roman-based writing system

Both

- Rich Morphological languages
- Read speech corpus from native speakers

Transcription task

- 1183 audio files (Total: 1h30) between 3 and 7s from both corpus
- Each file published as a HIT (USD 0.05)
- To avoid inept Turkers, HIT description and instructions were given in the respective languages
- For Amharic, we had the address of an online virtual keyboard

Completion rate

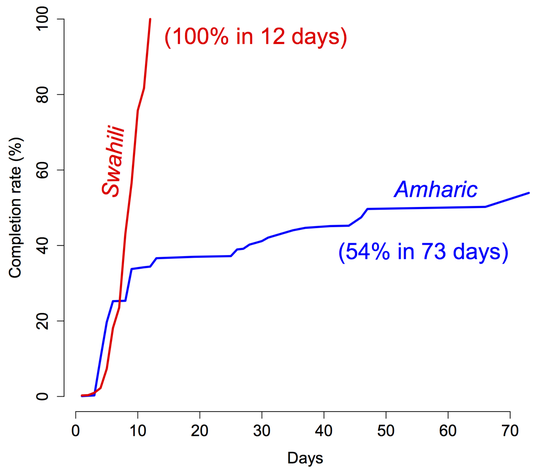

Figure 1: Completion rate per-day

Evaluation of Turkers transcriptions quality

Between reference transcriptions (REF) and Turkers transcriptions (TRK) at:

- Word-level / WER (standard)
- Syllable-level / SER (morphologically rich languages)
- Character-level / CER (some orthographic errors will not necessarily impact AM performances but will still inflate WER [4])

Table: *Error Rate (ER) of Turkers transcriptions*

	Amharic		Swahili	
Level	♯ Unit	ER (%)	♯ Unit	ER (%)
Wrd	4988	16.0	10998	27.7
Syl	21148	4.8	31233	10.8
Chr	42422	3.3	63171	6.1

Error Analysis

Most of the Swahili transcriptions were made by a second-language speaker

Table: *Most frequent confusion pairs for Swahili.*

Frq	REF	TRK	Frq	REF	TRK
15	serikali	serekali	6	nini	kwanini
13	kuwa	kwa	6	sababu	kwasababu
12	rais	raisi	6	suala	swala
11	hao	hawa	6	ufisadi	ofisadi
11	maiti	maiiti	5	dhidi	didi
9	ndio	ndiyo	5	fainali	finali
7	mkazi	mkasi	5	jaji	jadgi

- Wrong morphological segmentations
- Common spelling variations of words
- Misspellings due to English influence in loanwords
- Misspellings based on pronunciation
- Personal orthographic convention

Performance in ASR systems

- 3-gram LM (using SRILM) and 64k vocabulary
- 2 CI HMM-based Acoustic Model (using Sphinx): one learned with REF transcriptions and one with TRK transcriptions
- Phones: 36 for Swahili, 40 for Amharic

Table: *Performance of ASRs*

Languages	ASR	♯ Snt	♯ Wrd	WER
Swahili	REF	82	1380	38.0
	TRK	82	1380	38.5
Amharic	REF	359	4097	40.1
	TRK	359	4097	39.6

Conclusions

- Usability of Amazon's Mechanical Turk speech transcription for two under-resourced African languages.
- Similar AM's performance between REF and TRK transcriptions (even with 2nd-language speaker transcriptions)
- Not all languages are equal in completion rate.

English > > Swahili > > Amharic

Ethical issues

MTurk is proved to be powerful for NLP domains. However, It also happens to be controversial among the research community for legal and ethical issues. One should be careful on the manner the data are collected or the experiments are led.

- Systematically explain "who we are", "what we are doing" and "why" in HITs descriptions (as done traditionally for data collection);
- Make the data obtained available for free to the community;
- Set a reasonable payment so that the hourly rate is decent;
- Filter turkers by country of residence to avoid those who consider MTurk as their major source of funding.

For further informations

- Acknowledgment - Supported by the Pi ANR
- A full broadcast-news transcription process has been done for Swahili after this experiment but the choice has been made to work in collaboration with a Kenyan institute : The Taji Institute
- Contact - `hadrien.gelas@univ-lyon2.fr`, `solomon.abate@imag.fr`, `laurent.besacier@imag.fr`, `francois.pellegrino@univ-lyon2.fr`
- Full paper and poster can be found at: `http://www.ddl.ish-lyon.cnrs.fr/Gelas`

References

[1] A. Gruenstein, I. McGraw, and A. Sutherland, "A self-transcribing speech corpus: collecting continuous speech with an online educational game," in SLaTE Workshop, 2009.

[2] I. McGraw, A. Gruenstein, and A. Sutherland, "A self-labeling speech corpus: Collecting spoken words with an online educational game," in Interspeech, 2009.

[3] M. Marge, S. Banerjee, and A. Rudnicky, "Using the amazon mechanical turk to transcribe and annotate meeting speech for extractive summarization," in NAACL HLT 2010 Workshop.

[4] S. Novotney and C. Callison-Burch, "Cheap, fast and good enough: Automatic speech recognition with non-expert transcription," in NAACL HLT 2010.

[5] M. Marge, S. Banerjee, and A. Rudnicky, "Using the amazon mechanical turk for transcription of spoken language," in ICASSP, 2010.