Mining *Audio* and *Text* pairs for improving ASR Systems

AI4Bharat Speech Workshop

28 July 2022

How can we collect ***labelled*** data for training ASR systems?

What do we mean by labelled data?

- Audio-text pairs
- Sentence level data

How can we collect labelled data?

- Collect diverse data on ground
  - This is expensive
- Subtitle existing audio/video content (Eg. PlanetRead)
- **Mine data from existing content (Eg. NewsOnAIR)**
- Any other techniques?

Content in NewsOnAIR - I

Language	Stations	Bulletins	Hours
bengali	4	5729	646
gujarati	3	5597	683
kannada	3	4896	646
hindi	11	18764	2349
malayalam	3	5694	865
marathi	5	9020	1284
odia	2	5775	773
punjabi	1	889	122
sanskrit	1	731	64
tamil	4	7030	1118
telugu	4	5571	839
urdu	4	2884	306
All	**45**	**72580**	**9695**

Content in NewsOnAIR - II

- Audio data from different news bulletins and corresponding script used by the speaker

Audio Sample

Transcript

दिनांक 27.7.2022 समय – 02.20 च्ड

आकाषवाणी भोपाल से प्रस्तुत हैं प्रादेषिक समाचार

मिशन सेहत कार्यक्रम

प्रदेश में स्वास्थ्य सुविधाओं के विस्तार एवं दवाओं की उपलब्धता के लिए मिशन सेहत से जुड़े कार्यों के लिए 434 करोड़ रूपए का बजट निर्धारित किया गया है। ये जानकारी मुख्यमंत्री शिवराज सिंह चौहान ने कायाकल्प अवार्ड और मिशन सेहत कार्यक्रम की तैयारियों की समीक्षा करते हुए दी। इसमें से मरम्मत के कार्यों पर लगभग 146 करोड़ रूपए की राशि खर्च की जाएगी। इसके अलावा, रेडियोलॉजिकल जाँच की सुविधाएँ बढ़ाने पर 73 करोड़, पैथालॉजी जाँच सुविधाओं के लिए 119 करोड़, फर्नीचर और पलंग पर 35 करोड़ रूपए की राशि व्यय की जाएगी। अस्पतालों के लिए 20 हजार नए पलंग खरीदे जाएंगे। अस्पतालों में ब्लड बैंक एवं ब्लड स्टोरेज सुविधा के साथ डायलिसिस मशीनें भी बढ़ाई जाएंगी। समीक्षा बैठक में 8 अगस्त को इस संबंध में कार्यक्रम करने का निर्णय लिया गया।

.....

रुक जाना नहीं परीक्षा परिणाम

राज्य मुक्त स्कूल शिक्षा बोर्ड द्वारा रूक जाना नहीं योजना में कक्षा 10वीं और 12वीं का परीक्षा परिणाम कल घोषित कर दिया गया है। राज्य मुक्त स्कूल शिक्षा बोर्ड के निदेशक प्रभात राज तिवारी ने

Challenges with Audio Data

- Long intro and outro music at the start and end of the files.
- Non-transcribed segments like announcements, external news clips.
- Some files have background music for the entire duration.
- Misspoken words, fumbles

Challenges with Text data - I

- PDFs contain proprietary encodings (non-UTF8) due to legacy issues.
- Standard PDF parsers cannot be used to extract text from them

आकाशवाणी पोर्ट ब्लेयर
दिनांक : 27.07.2022
समय : 0705
<><><><><><><>

- करगिल विजय दिवस पर कल द्वीपों में कार्यक्रम आयोजित किए गए।
- दक्षिण अंडमान ज़िले से संबंध ज़िला नियोजन समिति की बैठक में कल वार्षिक कार्य योजना पर चर्चा की गई।
- आज़ादी का अमृत महोत्सव मनाए जाने के सिलसिले में कई कार्यक्रम आयोजित किए जा रहे है।
- तबादले पर दिल्ली रवाना होने से पूर्व मुख्य सचिव जितेन्द्र नारायण ने राज निवास में

03.04.2018 13.45 ஆல் இண்டியா ரேடியோ திருச்சிரப்பள்ளி - மாநிலச் செய்திகள்.

தலைப்புச் செய்திகள்

1. காவிரி மேலாண்மை வாரியத்தை உடனடியாக அமைக்க வலியுறுத்தி, தமிழகத்தில் அஇஅதிமுக சார்பில் இன்று உண்ணாவிரத போராட்டம் நடைபெறுகிறது.

2. ஊ.டீ.ளு.நு. பத்தாம் வகுப்பில் கணித பாடத்தில் மறுதேர்வு நடத்தப்படாது என்று, மத்திய பள்ளிக்கல்வித்துறை செயலர் அறிவித்திருக்கிறார்.

3. நாடாளுமன்றத்தில், அமளி காரணமாக இரு அவைகளும் 19-வது நாளாக இன்றும் அலுவல்கள் எதுவுமின்றி ஒத்திவைக்கப்பட்டன.

4. ளுஊஇ ளுவு வன்முறை தடுப்பு சட்ட மேல்முறையீட்டு மனு, இன்று பிற்பகல் 2 மணிக்கு உச்சநீதிமன்றத்தில் விசாரணைக்கு எடுத்துக்கொள்ளப்படுகிறது.

இனி விரிவான செய்திகள்

□□□□□

காவிரி மேலாண்மை வாரியம்

காவிரி மேலாண்மை வாரியத்தை உடனடியாக அமைக்க வலியுறுத்தி, தமிழகத்தில் அஇஅதிமுக சார்பில் உண்ணாவிரதப் போராட்டம் இன்று நடைபெற்று வருகிறது.

32 மாவட்ட தலைநகரங்களில் அஇஅதிமுக-வினர் உண்ணாவிரதம் மேற்கொண்டுள்ளனர்.

Challenges with Text data - II

MORNING, 07-04-2018, NATIONAL NEWS, 7.50 AM to 8.00 AM

1

પ્રસાર ભારતી

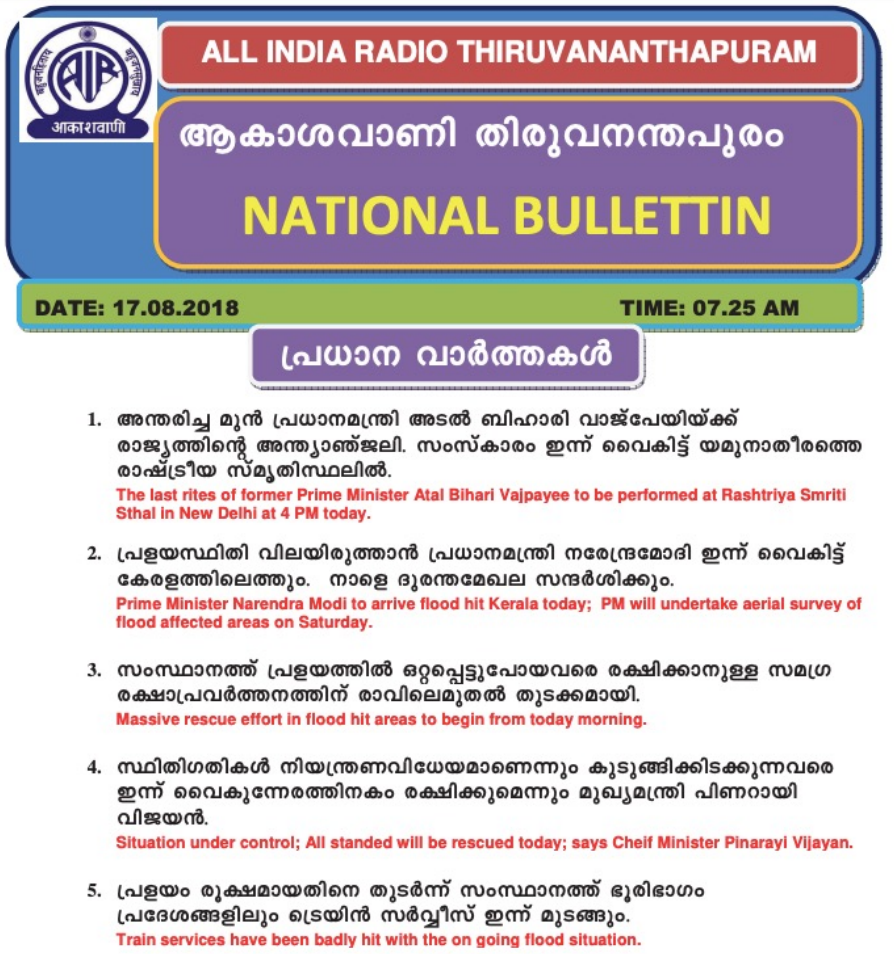
નેશનલ સમાચાર વિભાગ, આકાશવાણી, અમદાવાદ

Date : 07-04-2018 Morning :7.45 to 8.55

Day : SATURDAY National News

આકાશવાણી સમાચાર આશુતોષ રાવલ વાંચે છે.

- પ્રધાનમંત્રી નરેન્દ્ર મોદી આજે નવી દિલ્હીમાં નેપાળના પ્રધાનમંત્રી કે.પી.શર્મા ઓલી સાથે પ્રતિનિધિ મંડળ સ્તરની વાતચીત કરશે.
- રાષ્ટ્રપતિ રામનાથ કોવિંદ આજથી ત્રણ આફ્રિકી દેશો ગીની, સ્વાઝીલેન્ડ અને ઝાંમ્બીયાની મુલાકતે જવા રવાના થશે.
- સીરીયામાં વિદ્રોહીઓના અંતિમ ઠેકાણા એવા પુર્વીઘૌતાના એક શહેર ઉપર થયેલા હવાઈ હુમલામાં ચાલીસ નાગરીકોના મોત થયા છે.
- ગોલ્ડકોસ્ટ કોમન વેલ્થ રમતોમાં આજે ભારતની પુરૂષ હોકી ટીમ પોતાના અભિયાનની શરૂઆત કરશે.
- આજથી ઈન્ડિયન પ્રિમિયર લીગ ક્રિકેટ ટુર્નામેન્ટના અગીયારમાં સંસ્કરણની શરૂઆત થશે.



ALL INDIA RADIO THIRUVANANTHAPURAM

ആകാശവാണി തിരുവനന്തപുരം

NATIONAL BULLETTIN

DATE: 17.08.2018 TIME: 07.25 AM

പ്രധാന വാർത്തകൾ

1. അന്തരിച്ച മുൻ പ്രധാനമന്ത്രി അടൽ ബിഹാരി വാജ്പേയിയ്ക്ക് രാജ്യത്തിന്റെ അന്ത്യാഞ്ജലി. സംസ്കാരം ഇന്ന് വൈകിട്ട് യമുനാതീരത്തെ രാഷ്ട്രീയ സ്മൃതിസ്ഥലിൽ.
   The last rites of former Prime Minister Atal Bihari Vajpayee to be performed at Rashtriya Smriti Sthal in New Delhi at 4 PM today.
2. പ്രളയസ്ഥിതി വിലയിരുത്താൻ പ്രധാനമന്ത്രി നരേന്ദ്രമോദി ഇന്ന് വൈകിട്ട് കേരളത്തിലെത്തും. നാളെ ദുരന്തമേഖല സന്ദർശിക്കും.
   Prime Minister Narendra Modi to arrive flood hit Kerala today; PM will undertake aerial survey of flood affected areas on Saturday.
3. സംസ്ഥാനത്ത് പ്രളയത്തിൽ ഒറ്റപ്പെട്ടുപോയവരെ രക്ഷിക്കാനുള്ള സമഗ്ര രക്ഷാപ്രവർത്തനത്തിന് രാവിലെമുതൽ തുടക്കമായി.
   Massive rescue effort in flood hit areas to begin from today morning.
4. സ്ഥിതിഗതികൾ നിയന്ത്രണവിധേയമാണെന്നും കുടുങ്ങിക്കിടക്കുന്നവരെ ഇന്ന് വൈകുന്നേരത്തിനകം രക്ഷിക്കുമെന്നും മുഖ്യമന്ത്രി പിണറായി വിജയൻ.
   Situation under control; All standed will be rescued today; says Cheif Minister Pinarayi Vijayan.
5. പ്രളയം രൂക്ഷമായതിനെ തുടർന്ന് സംസ്ഥാനത്ത് ഭൂരിഭാഗം പ്രദേശങ്ങളിലും ട്രെയിൻ സർവ്വീസ് ഇന്ന് മുടങ്ങും.
   Train services have been badly hit with the on going flood situation.

Text Preprocessing Pipeline

- Use Google Document AI OCR [2] to extract text
- Remove characters that do not belong to the language.
- Split the text into sentences using End-of-Sentence (EOS) character detected by OCR.
  - Some sentences are not split as EOS is not detected by OCR
- Remove headings, non-spoken text at the start of the files based on heuristics.
  - Heading lines at start of file were generally less than 4 words long
- Convert numbers to words for better alignment.

[2] https://cloud.google.com/document-ai

Approach

Alignment of Audio and Text

News Bulletin *(Long Audio)*

प्रदेश में नवरात्रि पर्व की शुरूआत श्रद्धालु माँ दुर्गा की पूजा कर रहे है दुर्गा पंडलों में साजसज्जा की गई

Full Transcript

Audio-Text Alignment

प्रदेश में नवरात्रि पर्व की शुरूआत 

श्रद्धालु माँ दुर्गा की पूजा कर रहे हैं 

दुर्गा पंडलों में साजस ज्जा की गई 

Sentence-level *Aligned*

How can we solve the alignment problem?

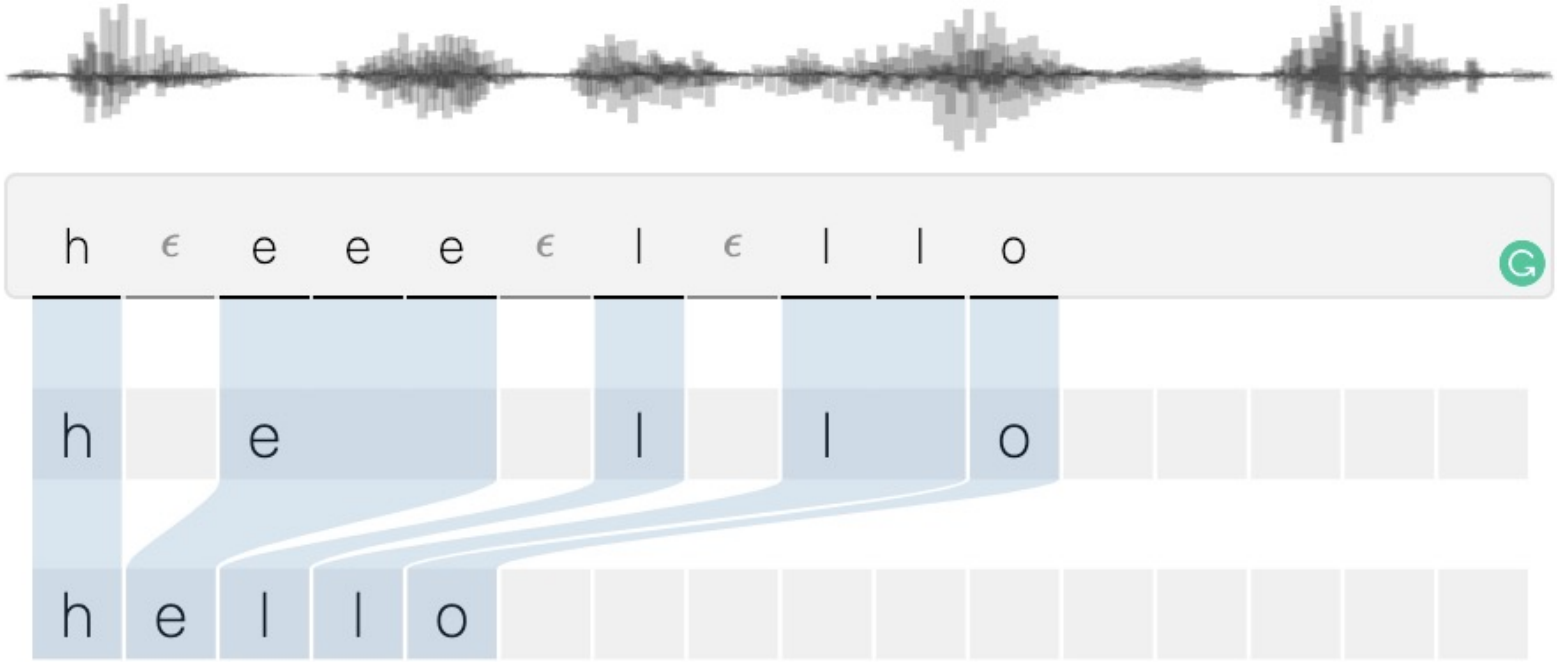
Alignment Process

- We follow a 3-step process for alignment -

1. Use existing ASR systems to generate text with timestamps (may be noisy)
2. Align noisy ASR text with reference text – Needleman Wunsch
3. Get sentence level audio-text pairs

1. ASR systems to generate text timestamps

- Utilize End-to-End ASR models trained using Connectionist Temporal Classification (CTC) [3] Loss to generate timestamps.



[3] https://distill.pub/2017/ctc/

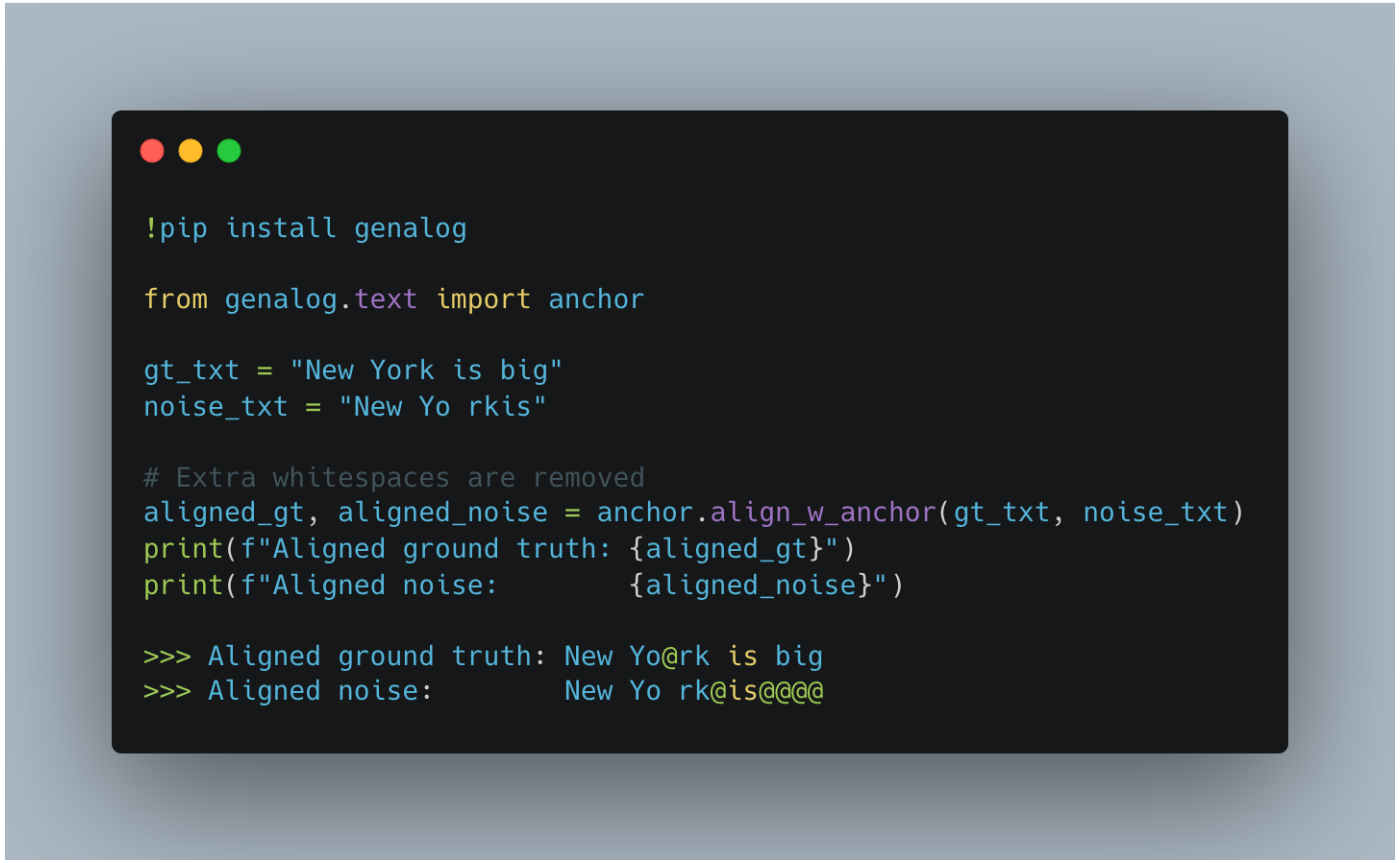
2. Align noisy ASR text with reference text

How to align reference text to ASR text?

Reference Text : "New York is big"

Noisy ASR output = "New Yo rkis"

Solution : Use Needleman Wunsch [4] Algorithm



```
!pip install genalog

from genalog.text import anchor

gt_txt = "New York is big"
noise_txt = "New Yo rkis"

# Extra whitespaces are removed
aligned_gt, aligned_noise = anchor.align_w_anchor(gt_txt, noise_txt)
print(f"Aligned ground truth: {aligned_gt}")
print(f"Aligned noise:        {aligned_noise}")

>>> Aligned ground truth: New Yo@rk is big
>>> Aligned noise:        New Yo rk@is@@@@
```

[4] https://en.wikipedia.org/wiki/Needleman%E2%80%93Wunsch_algorithm

3. How to get sentence level data?

- Story so far..
  - Reference Text is aligned to ASR Text
  - How to get sentence level audio-text pairs?
    - Find the start and end timestamp corresponding to the sentence boundaries.
  - Trim that part of the audio and the corresponding text for the sentence.

Results – #No. of Hours of Mined Data

6013 Hrs
Mined Data

/

9049 Hrs
Total Data

Results – Performance Improvement in ASR

18.80 WER
Existing data

14.16 WER
With Mined data

* Average WER across 7 Hindi Benchmarks

Future Directions

- Extend to all 40 Indian Languages available in NewsOnAIR
- Remove dependence on trained ASR systems to perform the alignment (unsupervised approaches)

Thank you

Questions?

https://ai4bharat.iitm.ac.in/