Konkani Integer Phonetic Transcription System

Swapnil Fadte [1] Edna Vaz Fernandes [23] Hanumant Redkar [1] Teja Kundaikar [1] Ramdas Karmali [1] Jyoti D. Pawar [1]

[1]Discipline of Computer Science and Technology, Goa Business School, Goa University, India [2]Department of Konkani, Govt. College of Arts Science and Commerce Quepem, Goa India [2]University of Mumbai, India

Objective of this Konkani Phonetic Transcription System is to built a resource that would phonetically transcribe integers in Konkani language and generate their corresponding written form in the Devanagari script

Abstract

In this work, we have built a resource that would phonetically transcribe Konkani Integers and generate their written form in the Devanagari script. The algorithm developed in this work made robust enough to automatically give a written form of any Konkani numeral in the Devanagari script, along with its equivalent IPA transcription. We have tried to keep the phonetic transcriptions as closer to their natural pronunciations. This is done for the purpose of capturing the general tendency of the language. So, for example, while the number '8' आठ [at^{h}] is written with an aspirated retroflex consonant ठ [ʈ^{h}], the final consonant [t^{h}] is heard without aspiration in the actual speech. This loss of aspiration at word final position generally happens across all the consonants of the language, in the Konkani varieties spoken in the state of Goa in India.

Introduction

India is a multilingual country having various languages and dialects. An Indian language - Konkani which belongs to Indo-Aryan language family, is an official language of the state of Goa in India. Anyone who understands a certain script can read any language written in that script. However, when it comes to the reading (pronunciation) of numerals, one needs to follow certain rules. Following are some representative examples in Konkani where we can see a combination of text with numerals. Such data instances in a text corpus could pose a big challenge for any system that aims to transcribe text accurately.

1. "ता. १ फेब्रुवारी २०२२" [ta. ek februvari don həd͡ʒar bavi:s] ('dtd. 1st February 2022'). The above character and numeral combination refers to a specific date and a year. While १ 'one' in १ फेब्रुवारी '1st February' might be easily transcribed as *ek (februvari)* by any system, the numeral 2022 which is a year, has to be read and transcribed as दोन हजार बावीस [don həd͡ʒar bavi:s] "two thousand twenty-two" and not as दोन शुन्य दोन दोन [don ʃunjə don don] "two zero two two".
2. "सकाळीं 08:00 ते 10:00 वरांमेरेन" [səkaɭĩ a:ʈ^{h} tɛ d^{h}a vərãmeren] '(from) morning 8 to 10 a.m.' (Lit. morning 8 to 10 hours till). This phrase specifies a certain time of the day. The system needs to acknowledge this context of time and generate a string that reads the numerals as hours (and minutes in some other temporalcontext).
3. "गोंयचें क्षेत्रफळ 3701 चौखण किलोमिटर आसा." [gõjt͡ʃẽ kʃetrə'fəɭ ti:n həd͡ʒar satʃɛ ek ʦɔʊk^{h}əɳ kilomiʈər asa] ('The (total) area of Goa is 3701 sq. kms.'). The numeral in the above sentence specifies the area of the region of Goa. The numbers should be read as one unit, i.e., as तीन हजार सातशे एक [ti:n həd͡ʒar satʃɛ ek] (Lit. "Three thousand Seven Hundred One") and not as individual numbers
4. "माशेलाचो पिन कोड 403 107." [maʃɛlaʦɔɔ pin'kod ʧar ʃunjə ti:n ek ʃunjə sa:t] ('The pin code of Marcela is 403 107'). Postal Index Number (PIN or simply PIN Code) refers to the six-digit number used by India Post in its postal code system. More commonly, the numbers indicating such a code are read by spelling out the numerals as discrete units.
5. "ताचो फोन नंबर 9850 403 107" [taʧɔ fon nəmbər nəʋ a:ʈ^{h} pã:ʦ ʃunjə ʧa:r ʃunjə tin e:k ʃunjə sa:t / taʧɔ fon nəmbər nain eʈ faiʋ d͡ʒiro fo:r d͡ʒiro t^{h}ri vən d͡ʒiro sɛvən]('His phone number is 9850 403 107'). Phone numbers can be read differently by different speakers. However, reading the numbers as discrete units is a good way to spell out the long number string.

The table 1 represents the position mapping rules that were used for phonetic transcription.

Sr No	Integer length	IPA	Devanagari Transcription
1	3	[ʃɛ]	शे
2	4	[həd͡ʒar]	हजार
3	6	[lak^{h}]	लाख
4	8	[koʈi]	कोटी
5	10	[ərəb]	अरब
6	12	[k^{h}ərəb]	खरब

Table 1. Position mapping rules.

Scope of the Work

We have made an effort to develop an automatic system for Konkani language that gives the phonetic as well as Devanagari transcription of a given integer. This is the first kind of work that aims to automatically transcribe Konkani numerals appearing in different contexts into the officially recognised Devanagari script along with the pronunciation of the numerals (in IPA).

Algorithms

Data: *integer*
Result: *transcription_text*
$y \leftarrow$ "";
$X \leftarrow input$;
$N \leftarrow len(X)$;
if $N \geq 12$ **then**
  $R \leftarrow assign\ last\ 11\ didgits$;
  $L \leftarrow X/10^{11}$; `/* removing last 11 digits from X */`
  $y \leftarrow left\_trans(L) + pos\_mapping(12) + right\_trans(R)$
else
  $y \leftarrow right\_trans(X)$;
end

Algorithm 1: Integer transcription.

Data: *integer*
Result: *transcription_text*
$y \leftarrow$ "";
$X \leftarrow input$;
$N \leftarrow len(X)$;
if $N \leq 2$ **then**
  if $N == 0$ **then**
    $y \leftarrow$ "";
  else
    $y \leftarrow int\_mapping()$; `/* Using table 1 transcription rule */`
  end
else
  if $N \leq 3$ **then**
    if $N == 100$ **then**
      $y \leftarrow int\_mapping()$
    else
      $y \leftarrow int\_mapping() + pos\_mappings(3) + right\_trans(R)$
    end
  else
    if $N \leq 5$ **then**
      $y \leftarrow int\_mapping() + pos\_mapping(4) + right\_trans(R)$
    else
      if $N \leq 7$ **then**
        $y \leftarrow int\_mapping() + pos\_mapping(6) + right\_trans(R)$
      else
        if $N \leq 9$ **then**
          $y \leftarrow int\_mapping() + pos\_mapping(8) + right\_trans(R)$
        else
          $y \leftarrow int\_mapping() + pos\_mapping(10) + right\_trans(R)$
        end
      end
    end
  end
end

Algorithm 2: right_transcription.

Data: *integer*
Result: *transcription_text*
$y \leftarrow$ "";
$X \leftarrow input$;
$N \leftarrow len(X)$;
if $N \geq 11$ **then**
  $R \leftarrow assign\ last\ 9\ digits$;
  $L \leftarrow X/10^{9}$; `/* removing last 9 digits from X`
  $y \leftarrow left\_trans(L) + pos\_mapping(10) + right\_trans(R)$
else
  $y \leftarrow right\_trans(X)$
end

Algorithm 3: left_transcription.

The Transcription System

This work presents an automatic phonetic transcription system for Konkani Integers. The system takes an integer as an input and generates its representation in word (the written form) along with its phonetic transcription. Figure 1 diagrammatically presents the transcription system.

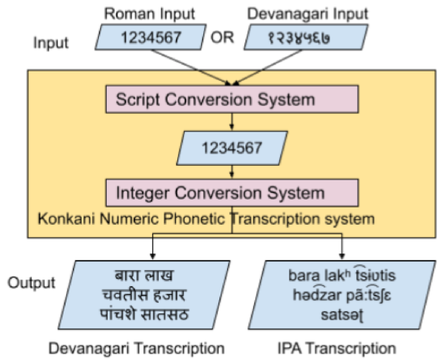

Figure 1. Integer Transcription System Diagram.

Methodology

The implementation of this transcription system consists of two parts. In the first part, the input is processed and standardised. In the second part, the standardised input is transcribed.

Transcription system converts the integer into its spoken form (its pronunciation in Konkani) by applying conversion rules and using a dictionary database for the numbers and their corresponding pronunciation. The algorithm for the transcription takes the integer and checks its length. Depending on the length of integer, it then calls for two sub algorithms: Left and Right transcription. Both Left and Right transcription works recursively and completes the transcription and results are merged and returned as final transcription of integers. Although it is rare to get an integer with a large number (length) in practice, ours algorithm can handle an integer of any length. Algorithms 1, 2 and 3 presents the implemented methodology.

Conclusion and Future Work

In this work, we have presented a system that transcribes Konkani integers into the officially recognised Devanagari script along with the IPA transcriptions of the numerals. The current system can technically handle integers of infinite length. As a future work, it can be extended for fractions, dates, scientific numbers, phone numbers, etc.

References

[1] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[2] G. K. Academy, *Konkani Shuddhalekhanache Nem 5th Edition*. Secretary,Goa Konkani Akademi 243, Patto Colony, Panaji , Goa - 403 001: Goa Konkani Academy, 1972.

[3] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, and et al., "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016.

[4] M. Junczys-Dowmunt, "Microsoft translator at WMT 2019: Towards large-scale document-level neural machine translation," in *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 225–233. [Online]. Available: https://aclanthology.org/W19-5321

[5] *The Constitution of India*, Eighth schedule, Article(s): 344(1) and 351, Description: Official languages, 1950.