



Vietnamese Speech Synthesis with End-to-End Model and Text Normalization

Do Tri Nhan, Nguyen Minh Tri, Cao Xuan Nam

Outline

01

Introduction

Overview about research

02

Related Work

Text-To-Speech
Tacotron2 and Waveglow

Method

Apply model for Vietnamese
Vinorm and Viphoneme

04

Experiment

Statistic and Result

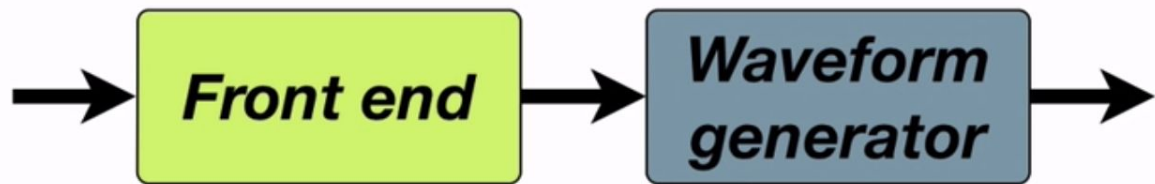
05



INTRODUCTION

Conventional Unit Selection

Unit Selection



text

"the cat sat"

*linguistic
specification*



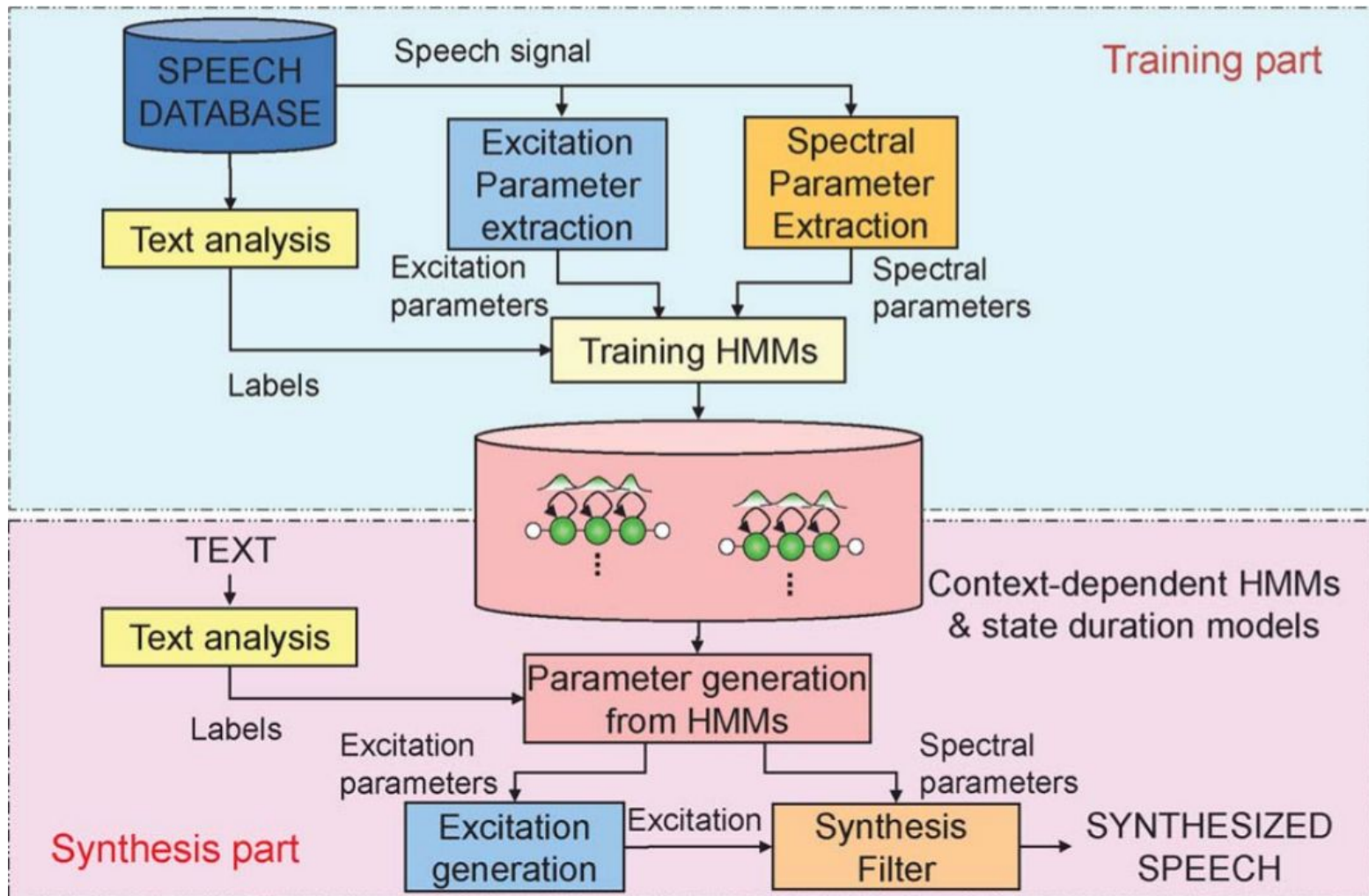
waveform



Unit Selection

- Knowledge based and Rule Based
- Speech units that match both **phonemes** and other **linguistic contexts** such as lexical stress, pitch accent, and part-of-speech information
- Speaking styles and emotions need larger speech databases

Hidden Markov Model



Statistic based model

- Data-driven approach: statistical parametric speech synthesis
- Acoustic parameters generated from HMMs selected according to the linguistic specification are used to drive a vocoder

Deep Neural Network

Some HMM OPEN SOURCE SOFTWARE TOOLS

- **Toolkit:**
 - HTS
 - HTK
 - Festival
 - hts_engine API
- **TTS system:**
 - Open JTalk
 - Flite+hts_engine

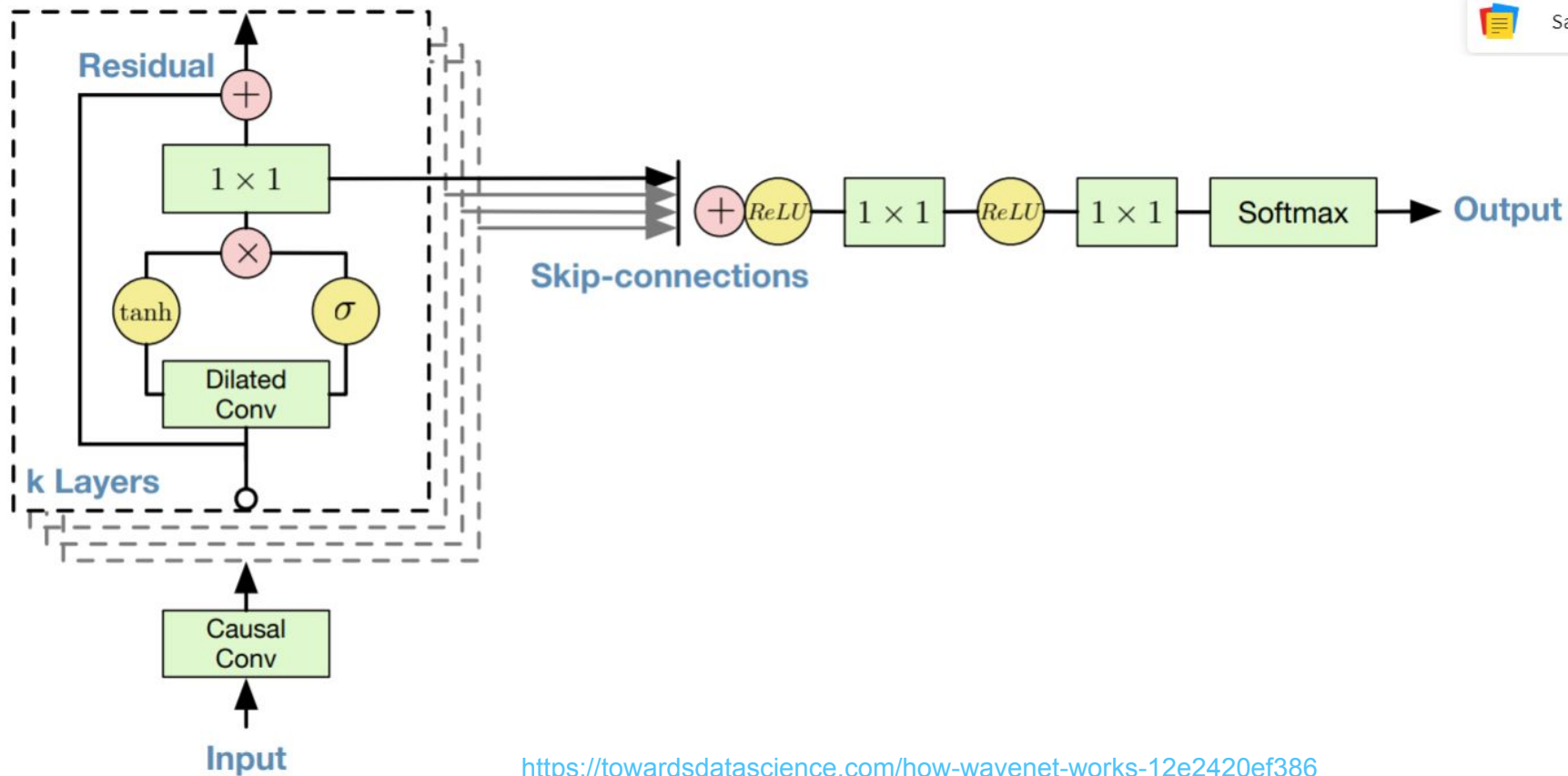
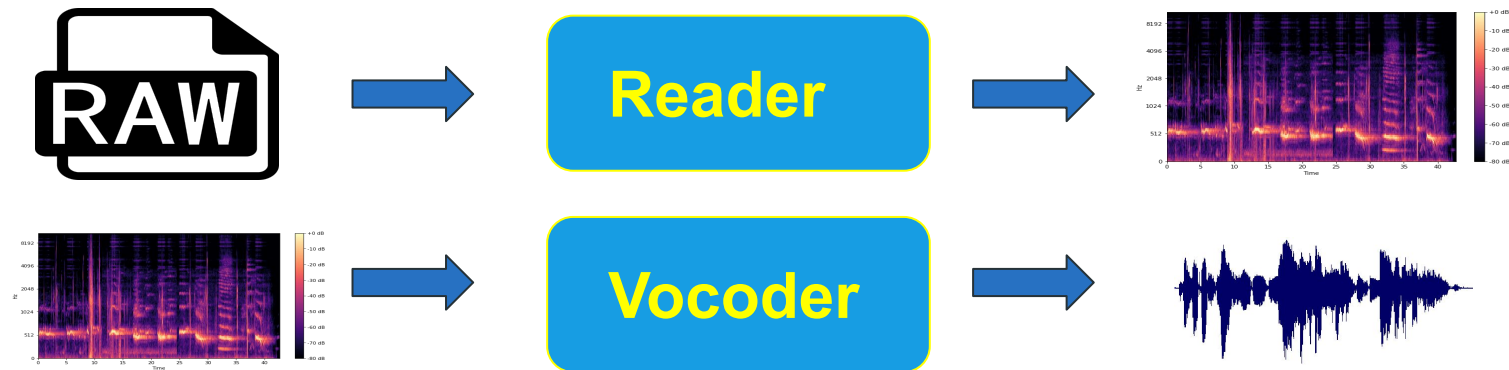


Figure 4: Overview of the residual block and the entire architecture.

End-to-End models



Learns to produce audio directly from text.

Some DNN speech synthesis model

Speech synthesis model	Mel Spectrogram Generation	Vocoder
Tacotron 2 (Google)	Attention based	Wavenet (2016)
Tacotron 1	Attention based	Griffin-Lim algorithm
Transformer TTS network		Wavenet (2016)
FastSpeech (Microsoft)	Feed-Forward Transformer	WaveGlow
DeepVoice3 (Baidu)	Attention based	All
Char2Wave	Attention based	SampleRNN (2017)
TTS Cube	Attention based	WaveRNN (2018)

Some DNN speech synthesis model

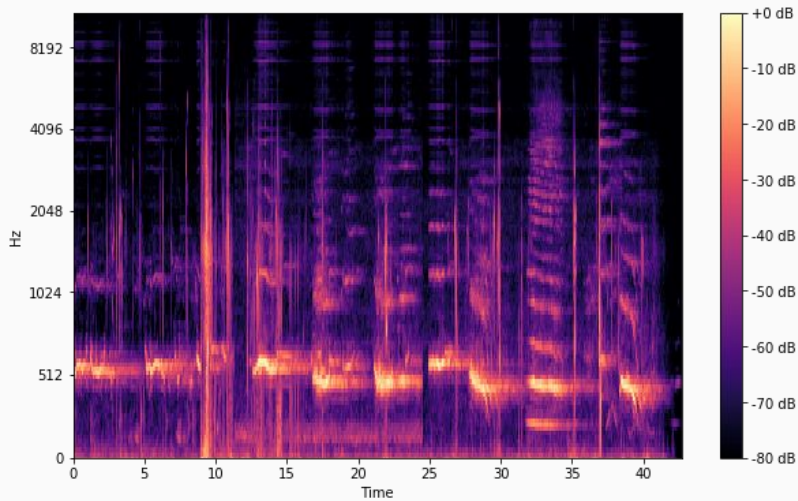
Method	MOS
<i>GT</i>	4.41 ± 0.08
<i>GT (Mel + WaveGlow)</i>	4.00 ± 0.09
<i>Tacotron 2 [20] (Mel + WaveGlow)</i>	3.86 ± 0.09
<i>Merlin [25] (WORLD)</i>	2.40 ± 0.13
<i>Transformer TTS [13] (Mel + WaveGlow)</i>	3.88 ± 0.09
<i>FastSpeech (Mel + WaveGlow)</i>	3.84 ± 0.08

Advantages

- Works well on **out-of-domain** and complex words
- Learns pronunciations based on **phrase semantics**
- Robust to spelling **errors**
- Sensitive to **punctuation**
- Stress on **capitalized** words
- Prosody changes when turning a statement into a **question**

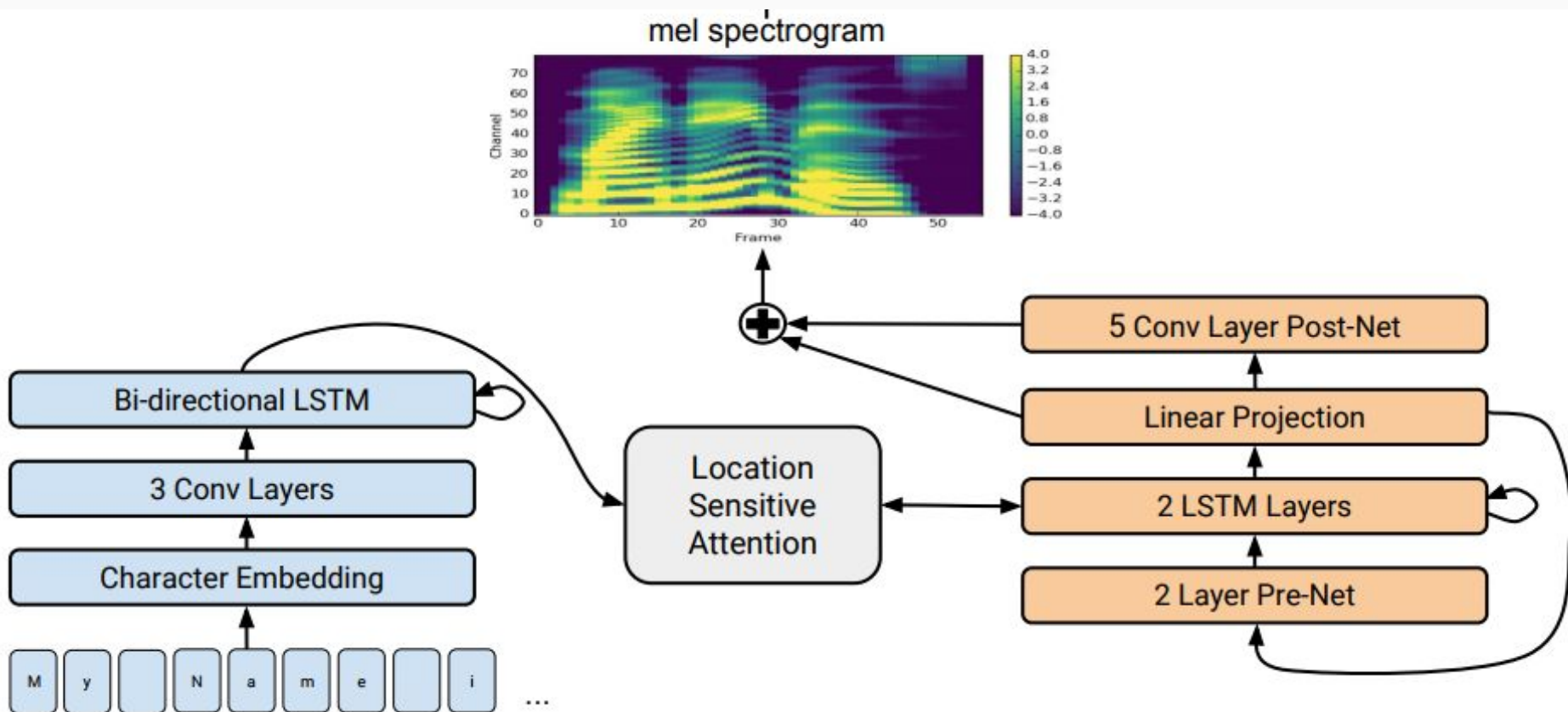


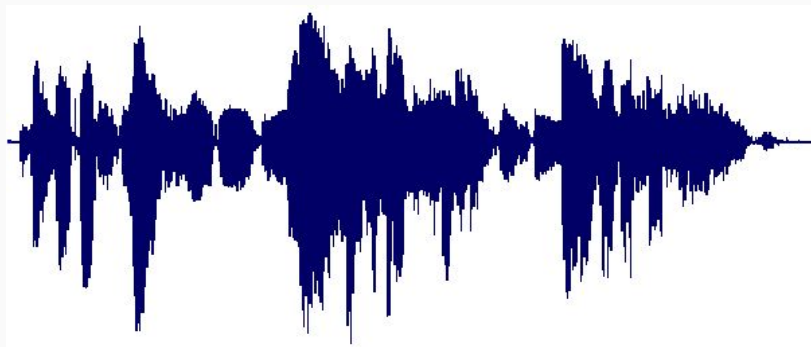
RELATED WORK



Generate a Mel-scale spectrogram

Attention based sequence-to-sequence model





Voice Synthesis

- Griffin-Lim algorithm (Tacotron)
- LSTM-RNN-based statistical parametric
- Wavenet-based

Speech samples	Subjective 5-scale MOS in naturalness	
	North American English	Mandarin Chinese
LSTM-RNN parametric	3.67 ± 0.098	3.79 ± 0.084
HMM-driven concatenative	3.86 ± 0.137	3.47 ± 0.108
WaveNet (L+F)	4.21 ± 0.081	4.08 ± 0.085

WaveGlow

- Sequential generation is too slow for production environments
- Method
 - **Flow based:**
 - Parallel WaveNet:
 - ClariNet
 - **WaveGlow**
 - FloWaveNet
 - **GAN**
 - **GAN-TTS**

Apply for Vietnamese?

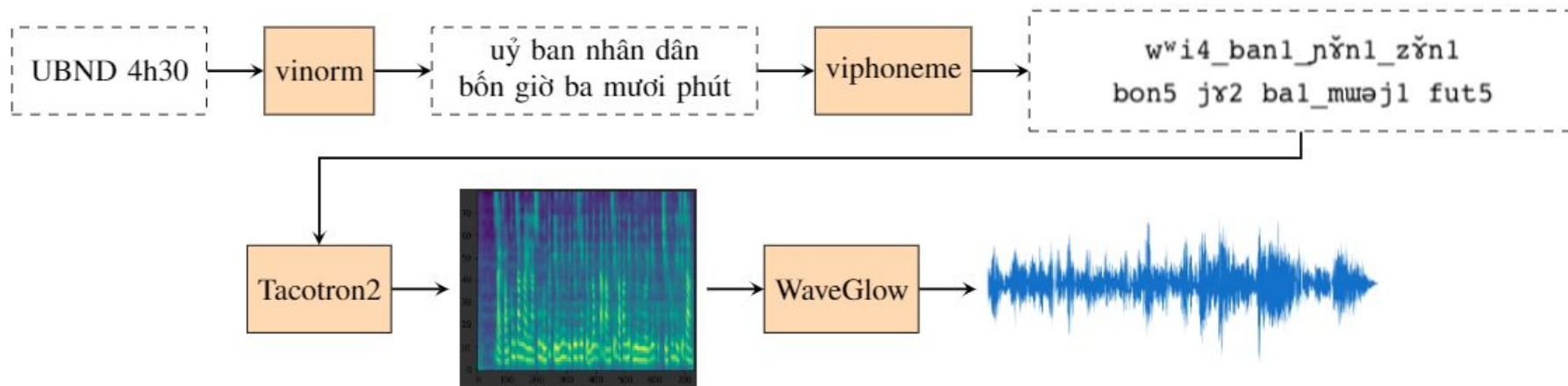
- USB wifi hãng **TP LINK**
- số điện thoại **1800.8098**
- trong khoảng thời gian **9h-12h30**
- tương đương 1n \approx 0,1kg)
- Khí **H₂S** còn gọi là khí Hidro Sunfua Hóa lỏng ở **60°C**
- Buôn Ma Thuột (**Đăk Lăk**)
- giá từ 500 triệu đến 1 tỷ **đồng/sào**
- cả Tấn Sinh (**1m83**) và Thành Chung (**1m80**) đều là mẫu trung vệ

Test cases from News are being used to improve Rules and expand Dictionary



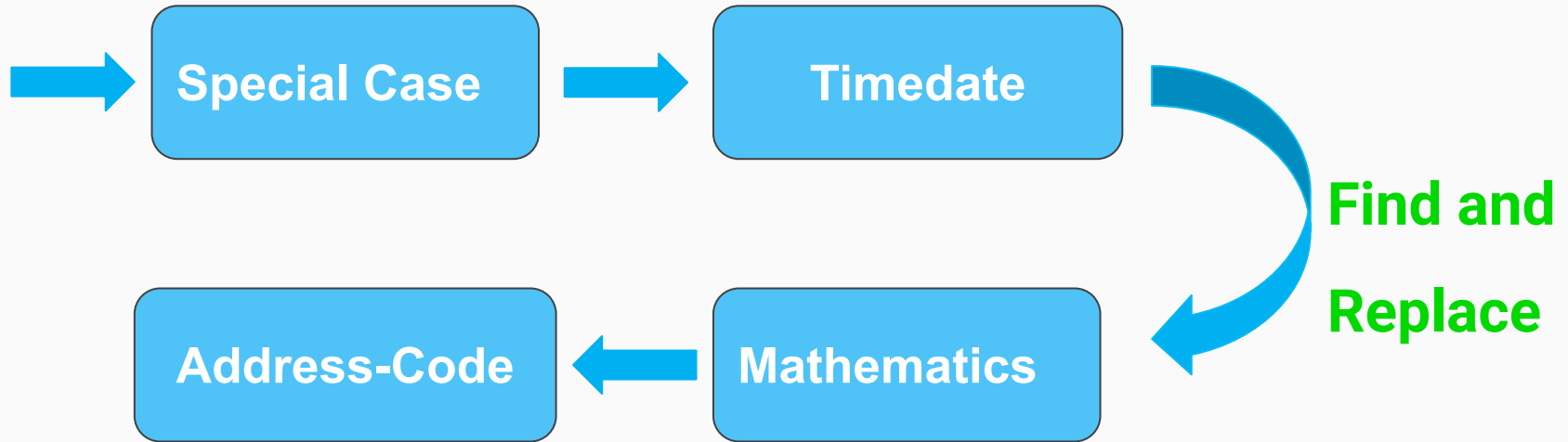
METHOD

Quick Overview



Vinorm

4 kind of rules



Ambiguous

Timedate

Classification

- **10/2019**
- **10.2019**
- **10-2019**
- **10,2019**

Address-Code

Mathematics

Mapping

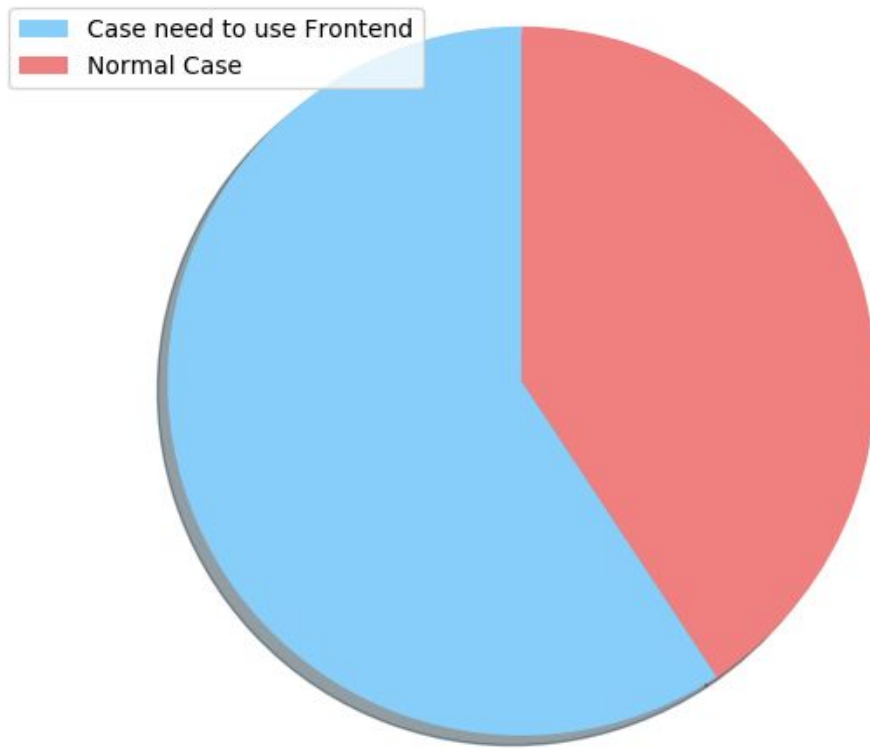
Abbreviations

**Teen Code - Slang -
Lingo**

**Symbol - Special
characters**

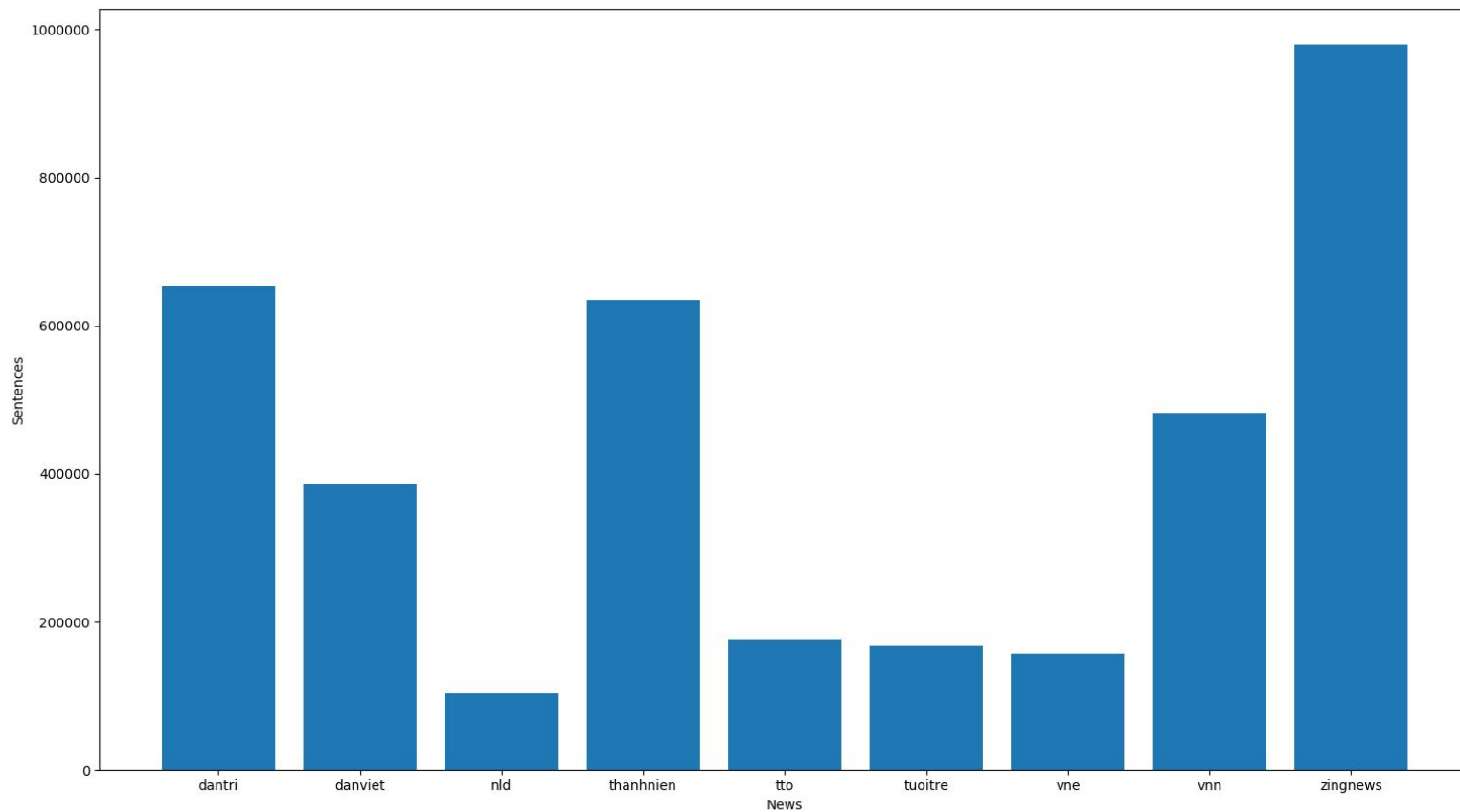


News



- **Total:** 6.308.173 sentences
- **Not Standardized:** 3.740.507 sentences

Sentences Distribution



[653545, 386444, 103218, 634974, 177287, 167162, 157202, 481566, 979109]

Random Test

60 tricky test cases

To compare TTS systems

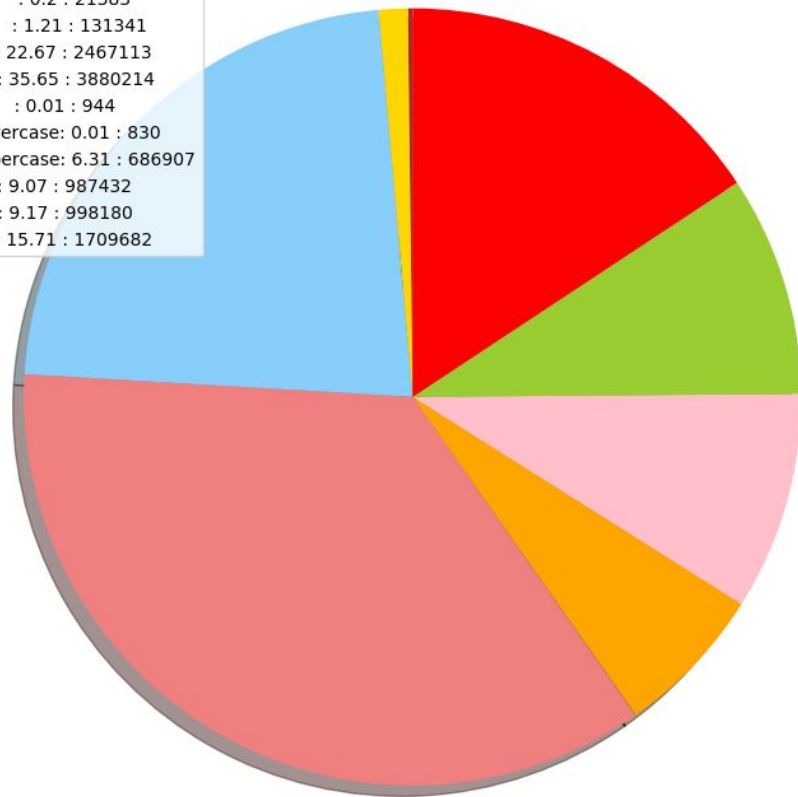
100 need-normalized case

Use as baseline to improve

500 random in practical contexts

Random test

Special Case	: 0.2 : 21383
Timedate	: 1.21 : 131341
Math	: 22.67 : 2467113
English	: 35.65 : 3880214
Teencode	: 0.01 : 944
Acronyms Lowercase	: 0.01 : 830
Acronyms Uppercase	: 6.31 : 686907
Initialism	: 9.07 : 987432
Proper	: 9.17 : 998180
Other	: 15.71 : 1709682



- Special case: website, phone, football, email
- Math: measure, digit, roman
- Initialism: total uppercase but not in English or Acronyms
- Proper: Uppercase first letter, not in dictionary

Insights

- Special Case is just 0.2%
- Most of the Timedate cases are the publication date of the news
- English accounts for 40 percent
- Acronyms Upper and Lower problem

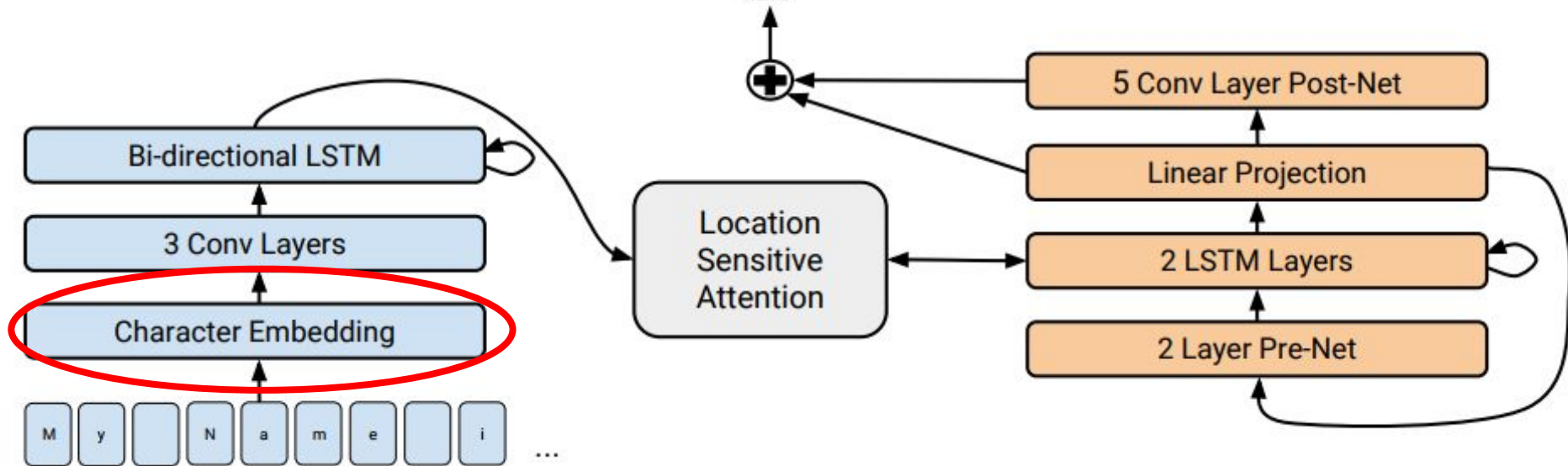
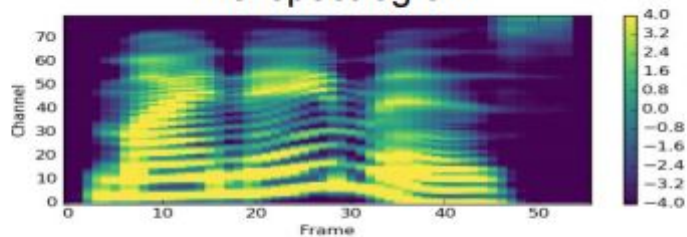
**English accounts for 40% of sentences
need to be normalized.**

Viphoneme

Text Phonetizer

Character Embedding to Phoneme Embedding

mel spectrogram



How Amazon Polly works with these case:

- Particular words, such as company names,
- Acronyms
- Foreign words
- Neologisms (e.g., “ROTFL”, “C’est la vie”)

Use custom lexicons to customize the pronunciation of **Nguyen**

```
<lexeme>
  <grapheme>Nguyen</grapheme>
  <grapheme>nguyen</grapheme>
  <grapheme>NGUYEN</grapheme>
  <phoneme>"nu.jEn"</phoneme>
</lexeme>
```

International Phonetic Alphabet (IPA)

```
<speak>
  You say, <phoneme alphabet="ipa" ph="pɪ'kɑ:n">pecan</phoneme>.
  I say, <phoneme alphabet="ipa" ph="'pi.kæn">pecan</phoneme>.
</speak>
```

Extended Speech Assessment Methods Phonetic Alphabet (X-SAMPA)

```
<speak>
  You say, <phoneme alphabet='x-sampa' ph='pI"kA:n'>pecan</phoneme>.
  I say, <phoneme alphabet='x-sampa' ph='""pi.k{n'>pecan</phoneme>.
</speak>
```


Cross Language (English)

- Popular English:
 - Singapore->sin ga bo
- With other words: (have meaning and available in Language Dictionary)



This is not totally standards

Understand -> **ən.dʒɪːˈstænd**

Text Phonetizer

$$(C_1)(w)V(G|C_2)+T$$

- C_1 = initial consonant onset
- w = labiovelar on-glide /w/
- V = vowel nucleus
- G = off-glide coda (/j/ or /w/)
- C_2 = final consonant coda
- T = tone.

Text Phonetizer

- Special case: Gì, Quyết (**Qu + uyể + t**)
- Tones for Vietnamese: blanks, grave, acute, hook, tilde, dot accents with the numbers from 1 to 6
- English phoneme: ["dʒ", "ð", "θ", "ʃ"]

Text Phonetizer

144 grapheme

['ʷəj], 'řj], 'wiə], 'řw], 'ʷəw], 'wet], 'iəw], 'uəj], 'wen], 'tʰw], 'wř], 'wiu], 'kwi],
'ŋm], 'k̠p], 'cw], 'jw], 'uə], 'eə], 'bw], 'oj], 'wi], 'vw], 'ăw], 't̪w], 'ʃw], 'aʊ], 'fw], 'ɛu],
'tʰ], 't̪], 'ɔɪ], 'xw], 'wɾ], 'ř], 'ɲw], 'ʊə], 'zi], 'wă], 'dw], 'eɪ], 'aɪ], 'ew], 'iə], 'ɣw],
'zw], 'ʷj], 'wɛ], 'ʷw], 'ɾj], 'ɔ:], 'əʊ], 'wɑ], 'mw], 'ɑ:], 'hw], 'ɔj], 'uj], 'lw], 'ɪə], 'ăj],
'u:], 'aw], 'ɛj], 'iw], 'aj], 'ɜ:], 'kw], 'nw], 't̪], 'ɲw], 'eo], 'sw], 'tw], 'z̪w], 'iɛ], 'we], 'i:],
'ʷə], 'd̪], 'ɲ], 'θ], 'ʌ], 'l], 'w], 'l], 'ɪ], 'ʷ], 'd], 'f], 'p], 'ə], 'u], 'o], 'ɜ], 'ɣ], '!', 'ð], 't̪],
'ɔ], 'ɜ], 'z̪], 'z], 'v], 'g], 'ă], '_, 'æ], 'ɾ], '2], 'd̪], 'i], '_, 'p], 'b], 'h], 'n], 'ʃ], 'ɔ], 'ɛ], 'k],
'm], '5], '_, 'c], 'j], 'x], 't̪], '_, '4], 'ʊ], 's], 'ɲ], 'ɑ], 'f], '?], 'r], '_, 'ɲ], 'f], '_, 'e], 't], '"]

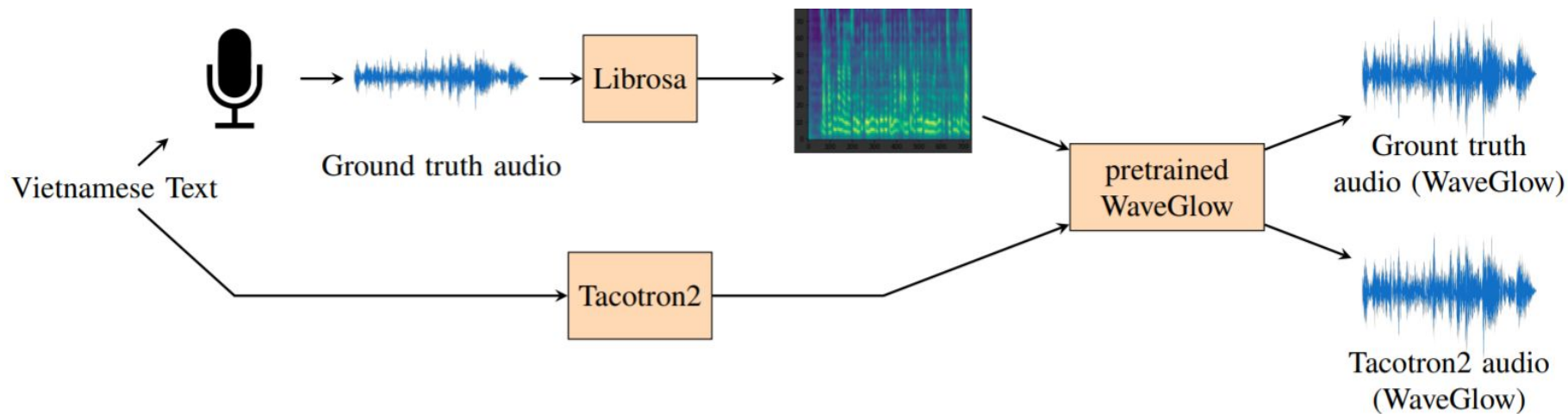


EXPERIMENT

Dataset and Training

- 240 hours for training
- BigCorpus VLSP 2019 (22 hours, 13462 utterances)
- Quadro K6000
- Audio Sample rate: 22050 Hz
- NoiseReducer

Evaluation



Mean opinion score

Model	MOS
Tacotron2 (WaveGlow)	3.97
Groundtruth (Mel + WaveGlow)	4.43

- Range from **1 to 5**, collect from **20** people

Viphoneme

Mel cepstral distortion measure

$$\frac{10\sqrt{2}}{\ln 10} \frac{1}{T} \sum_{t=1}^T \sqrt{\sum_i \left(c_{ti} - \hat{c}_{ti} \right)^2}$$

691,9 average from **200** audio tests

- Measure of how different two sequences of mel cepstra are

- **Vinorm:** <https://pypi.org/project/vinorm/>

```
from vinorm import TTSnorm
S=TTSnorm("Chuẩn hóa hệ thống nói tự động, gồm RegEx và Acronyms")
```

- **Viphoneme:** <https://pypi.org/project/viphoneme/>

```
from viphoneme import vi2IPA
phoneme = vi2IPA("Được viết vào 6/4/2020, có thể xử lý những trường hợp chứa  
>> dwək6 viət5 vaw2 ɣăw5 tʰaŋ5 bon5 nă̌m1 haj1 ɲin2 xoŋ̌m1 t̚ă̌m1 haj1_mwəj1 , kɔː
```

- **Appendix:** https://github.com/NoahDrisort/NICS_Appendix

- **Audio Sample:** <http://mostts.herokuapp.com/>

THANKS

AILAB and research funding from Advanced
Program in Computer Science, University of Science
Vietnam National University - Ho Chi Minh City

Special Case

Special Case

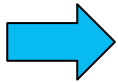
Telephone

- Must have 0 or + d {1,3} at the beginning and have formats like
 - International standard phone number:
 - +1 876 9 248 212 +972 3 52 23 161 +800 86 100 999
 - Vietnamese phone number pattern:
 - 0989.72.77.99 +84 989.72.77.99
 - USA phone number pattern:
 - +353 1 679 3958
- Hotline:
 - 19001288 1800.1288

Special Case

Website

- Prefix https ftp hoặc www
 - [www.google.st](#) | not match: [http://.com](#)
- Suffix is popular domain:
 - Com|au|.uk|co|.in|net|org|info|coop|int|co|.uk|org|.uk|ac|.uk|uk
 - [docs.google.com/document](#)



Pronounce each letter and character, ensuring integrity, don't read letter to sound like old VOS version

The word "com" reads the whole word to make it more natural

Special Case

Email

- Use common email rules
 - Spell each letter and symbol in English
 - “@” is read as symbol “a còng”
 - Dot and Slash: is sounded : “chấm” and “xuyệt”
 - If contain “gmail.com” => replace as “giy meo chấm com”
- Example
 - tripx.vn@gmail.com

Special Case

Football

Some number pattern:

- Đội hình 4-2-3-2
- Tỷ số 12-0
- Hạ Thái 4-3
- U.22

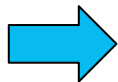
 Hyphen-minus and Dot will not be spelled

Date & Time

Time & Date

Time

- Suffix: am/pm
 - 12:34 AM 12:00 AM - 1:00 PM
- Signal of time: h/g
 - 12h 12h30 12h30 - 13h
- The validity of time
 - 22:30 but not 40:60



If "-" followed by captured regex will be read as “đến”

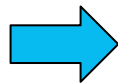
Add “giò” to suitable places

Time & Date

Date

- Typical pattern of the date month year
 - 5/12/2019 05.12.2019 5-12-2019
- Contain roman number
 - IV-2019 III/2019
- Prefix is a word of time
 - Ngày|sáng|trưa...chiều|tối|đêm|hôm...
 - Hôm 5.12 sáng 5.12 - 6.12

Add “ngày” “tháng” “năm” to suitable positions



Check the validity of date and time -> not 20-20-60

Time & Date

Date

- Month
 - Tháng 12/2019
- Date from to
 - Từ 5-31/12
 - Ngày 5/12 - 31/12
- Month from to
 - Từ 7-12/2019
 - Tháng 7.2019 - 9.2019

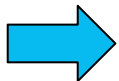
Mathematics

Mathematics

Floating Point | Normal Number

- Floating point is comma
 - 10,5 -10,5
 - 1.000.000,5 1.324.599
- Floating point is dot
 - 1,000,000 1200,000,000 1209,012,375,558
 - 12.5 12

Floating point is replaced by “phẩy”



Integer Part is read as Normal Number and discard frontier zero

Fractional Part: 2,05 => hai phẩy **không** năm

Mathematics

Unit of Measurement

- Unit is alphabetic:
 - Normal Number + {Alphabetic}
 - Normal Number + {Alphabetic} / {Alphabetic}
 - ➡ Check if match {Alphabetic} is in BaseUnit Mapping
- Unit is symbol
 - Normal Number + symbol
 - ➡ Check if match “symbol” is in CurrencyUnit Mapping

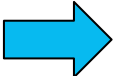
Mathematics

- The “/” is replaced by “trên”
- If {Alphabetic} is not in Dict => return the origin matching
- Distinguish upper and lower case Unit: **kb** and **kB**

Example:

- 100 kg/ha  kg and ha in Dict  100 ki lô gam trên héc ta
- 160.000-180.000 đồng/kg.

The Limit:

- 12A  12 am pe **OR** (lớp) 12 a
=> Context for unit

Mathematics

Roman Number

- Just consider Uppercase is able to be Roman number:
 - Lần thứ VI **AND** lần thứ vi
- Consider alphabet of Roman is [X,I,V],
 - XVII **AND** MV (1005)
- Check valid:
 - romanA -> decimal -> romanB => compare A=B?
 - IVX

Address - Code

Address

Political Division

Street

Office

- Kp|q|p|h|tx|tp|x
 - KP.3 H.Bình Chánh Q.7
- đường|số|số nhà|nhà|địa chỉ|tọa
lạc|xã|thôn|ấp|khu phố|căn hộ|cư xá|Đ/c
 - Căn hộ C03.09 ấp 5A
 - “/” is replaced by “xuyệt”
- phòng||lớp|đơn vị
 - Lớp 12a3 phòng F203
 - “/” is not read

Address

Codenumber

- Match All other cases contain numbers
- Some cases to deal with
 - **12A-7** : “-” is not spelled
 - **B007** : read each zero
 - **S-600** : Read continuous number string
 - **làn/m2** : In case of there is continuous letter string in dict, keep it
 - **C100045**: If length of continuous number string > 4 , spell each number
 - **TIM9+** : If totally Uppercase, spell each letter
 - **6λ** : Case includes special character | symbol
 - **.l86,** : Trim character

Update Dictionary

After run all rules, String just contain letters, space and special character

Run over each token to check if it is readable then mapping if necessary

– Dictionary: Vietnamese syllable

An orange rounded rectangle with a thin black border, containing the word "Popular" in white text.

Popular

- Includes **7698** syllable
 - Words that the current backend can support
 - The future can be extended with further
 - **Đăk Lăk** does not exist
-

— Mapping

Abbreviations

Teen Code - Slang - Lingo

Symbol - Special characters

Abbreviations

- **Acronyms:** Mapping with origin phrase
 - NSUT GDDT LD-TB&XH
- **Initialisms:**
 - STEM UNESCO
- **Spell each letter (not phonemes)**
 - SJC, PNJ, VNG, FBI, AFF
 - Skip it in this step and consider as unknown

Build Initialisms from:

- Total uppercase words
- Not appear in Acronyms Dict and not be an acronyms (case: some acronyms have not listed in Dict yet)
- Appears many times. **Threshold Frequency** is in discuss

The way Google read Acronyms

- **Uppercase:** replace with origin phrase
- **Lowercase:** read as unknown words

bhyt vs BHYT

ktx vs KTX

=> Limit misunderstandings when not sure

Teencode - Slang - Lingo

- Add more than 300 Slang to Teencode Mapping (366)

H'Hen Niê

H'Mông

- Total Uppercase Teencode will not be mapped:

BYT

TRC

- Add some Lingo that not appear in Popular Dict (Backend can not read)

Đăk -> Đắc

Lăk -> Lắc

- Difference style for the same word:

Hoá - Hóa

thuỷ - thủy

tuỳ - tùy

Symbol - Special Character

- Split letters and symbols

Ex: đi &chơi (check rule) => đi & chơi (check dict) => đi và chơi

- Punctuation

- “.” “!” “.” “?”
- “,” “;” “/”

- Check symbol

- ε (ép si lon) ε (thuộc) ζ (de ta)
 - η (ê ta) θ (thê ta) ι (i ô ta)
-

Tokenize Problem

- Tokenize with special character

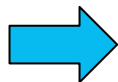
GD-ĐT

NQ/TW

- Tokenize with space

Ê-kíp

Ra-đi-ô



Combine two way: Tokenize with space first, if the token is unknown, then tokenize with special character

Unknown Token

Not appear in any
Dictionaries,
English
Proper Noun

The way Google read unknown word

- **Uppercase:** Spell each letter in **English**
- **Lowercase:**
 - Include vowel: -> keep ->letter to sound
 - Not Include: Spell each letter in English (sometimes Vietnamese)

bytero vs BYTERO (letter to sound và spell each letter)

ubndta vs UBNDTA (VI vs EN spell)

khtnm vs KHTNM (EN vs EN spell)

Apply for VOS

- **Lowercase:**

- If containing vowels: Not change | cut down
- Not containing vowels: Spell each as Vietnamese letter

- **Uppercase:**

- Spell each as English letter (**not phonemes**)

MVC => em vi si **not** mờ vờ cờ

Experiment

Programming Languages

Language Comparison:

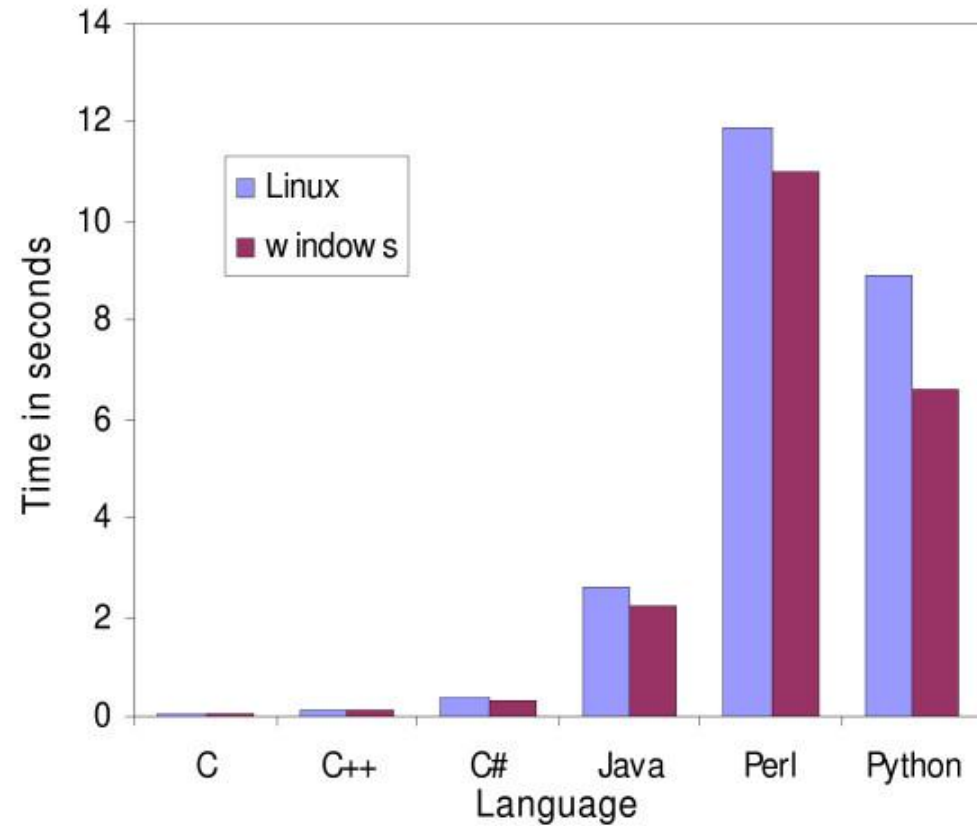
C++

Python

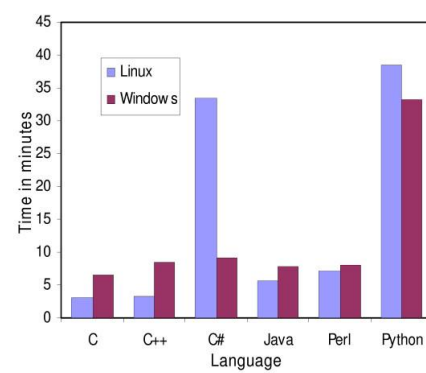
Perl

A comparison of common programming languages used in bioinformatics

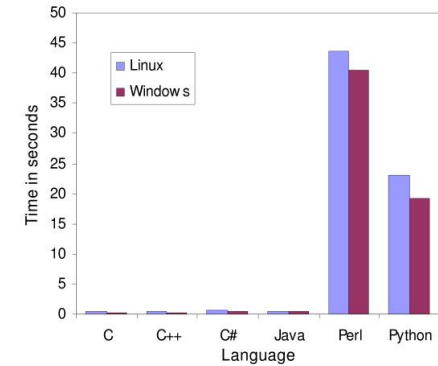
Holds the DNA **sequences** in memory, performs different computing tasks on the sequences (**text mining** or **text parsing**)



Neighbor-Joining tree construction algorithm



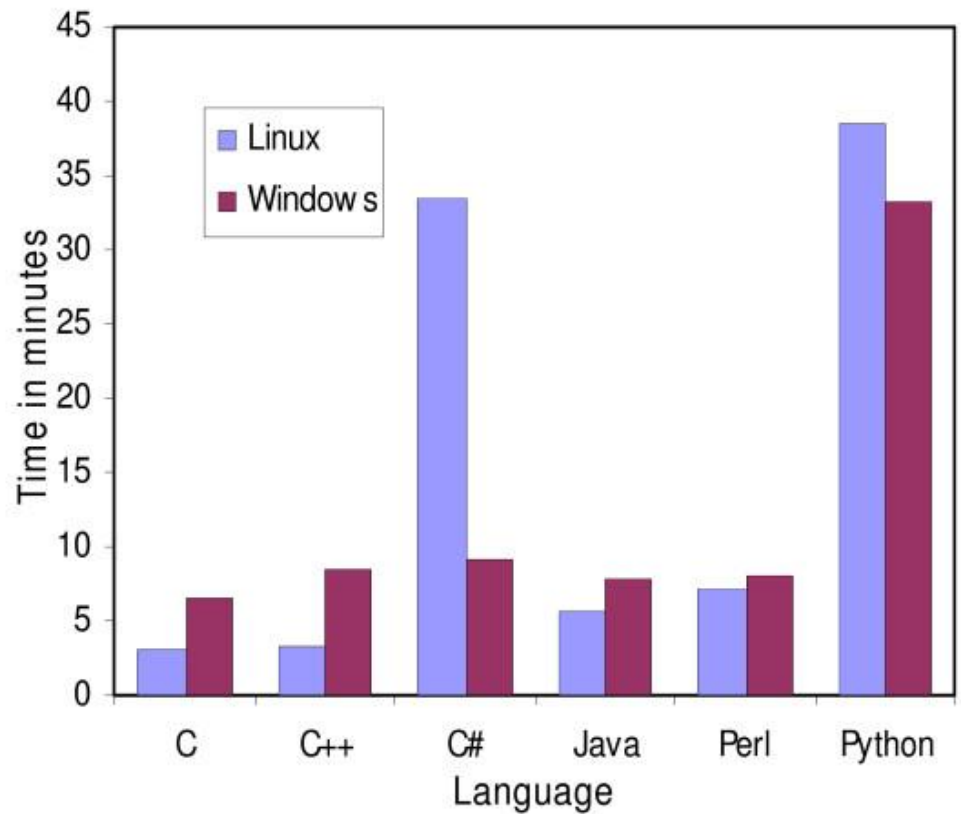
BLAST parsing



Sellers algorithm

Similar with TTS front-end

Compare to **C**, when **C++** standard libraries (ie. **character strings**) were used, the performances tended to **slightly** deteriorate



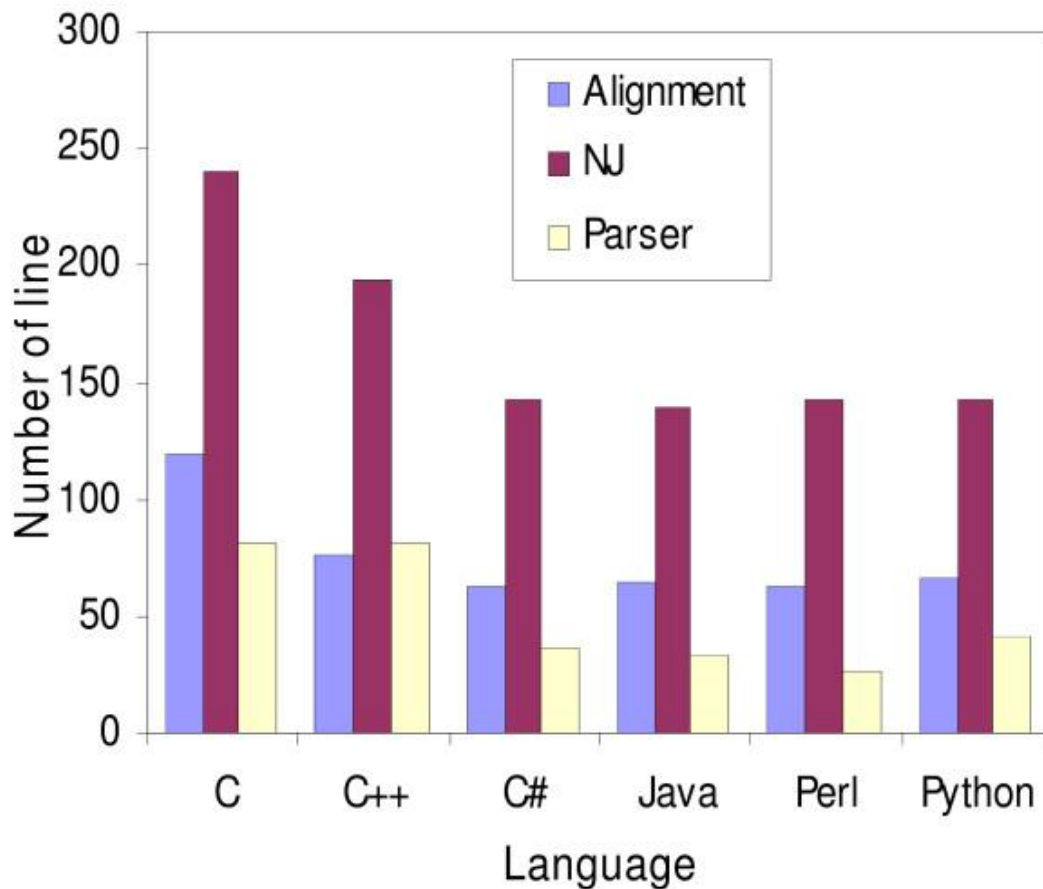
Tokenization was twice as fast as regular expressions for parsing the same BLAST file, but it took more time to write the program using tokens

BLAST parsing

Number of lines for each algorithm

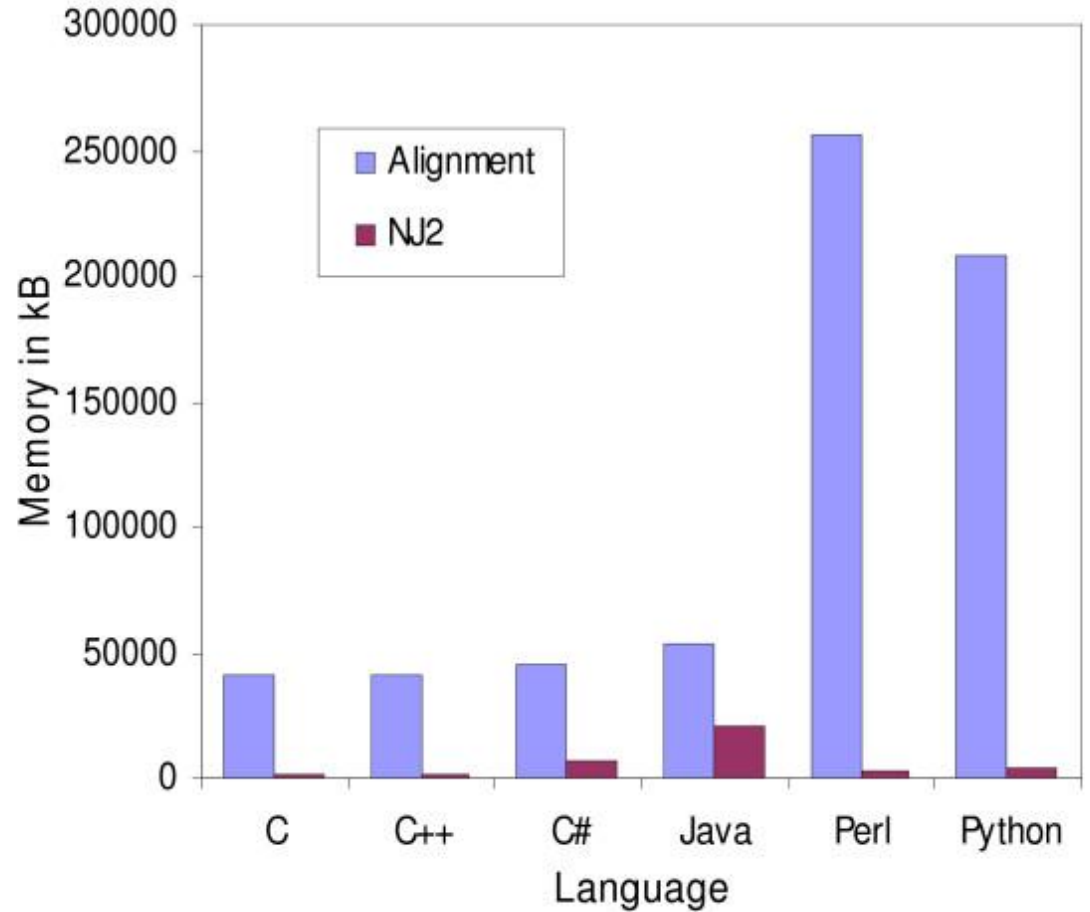
Perl : a unique statement can be used to detect a pattern

C has the highest number line of code, especially in Parser



Memory usage

Perl is use for string manipulation but it need **too much space** to store the sequences in memory



Usage

- **Python and Perl** are often called script languages and suitable for **web scripting, parsing**
 - **C and C++** are fully compiled languages, suitable for **system-intensive tasks.**
-

Programing Language of some TTS system

- **Festival** Speech Synthesis System of U.Edinburgh in **C++** less portable ,more robust,
- **Flite** (festival-lite) of CMU: **C** library but it can be used from a **C++**
- **Hts_engine** use for Jtalk, Sinsy written in **C++ and C**
- **eSpeak** written in **C**
- **ResponsiveVoice** (Web Speech API Specification) in **Javascript**

Language: C++

- Cross-Platform Deployment
- Fast running time
- Reuse the source code of VOS 2.0
- Not support Unicode natively (could use ICULIB instead)
- Steeper learning curve compared to Python
- Use less memory than Perl

Libraries

- ICU4C (version 64.2)
 - Standard Template Library
-

ICU4C Library

- Opensource.
 - Well-documented, robust and reliable.
 - Supporting Unicode String (including Vietnamese).
 - Regular Expression for Unicode.
 - Basic regular expression operators
 - Case Insensitive Matching and other flags
-

Project Structure

Folder

- Dict/
 - Dictionary files
 - Mapping/
 - Mapping files
 - RegexRule/
 - Regular Expression Rules
-

Files

- Input.txt
 - Regular text
 - Output.txt
 - Normalized text
 - Main.cpp
 - Main processing
 - SpecialCase.*, DateTime.*, ...
 - Rule Processing
 - ICUMapping.*, ICUReadFile.*, ...
 - Helper functions using ICU4C
-

Rule Processing

```
void loadPatterns(...) {  
    // Load Regex from files  
}
```

```
UnicodeString normalizeText(...) {  
    // Process regular text and return normalized text  
    // using this rule  
}
```

```
UnicodeString stringForReplace(...) {  
    // Processing the text matched by regex.  
}
```

Test and Statistics

News Corpus

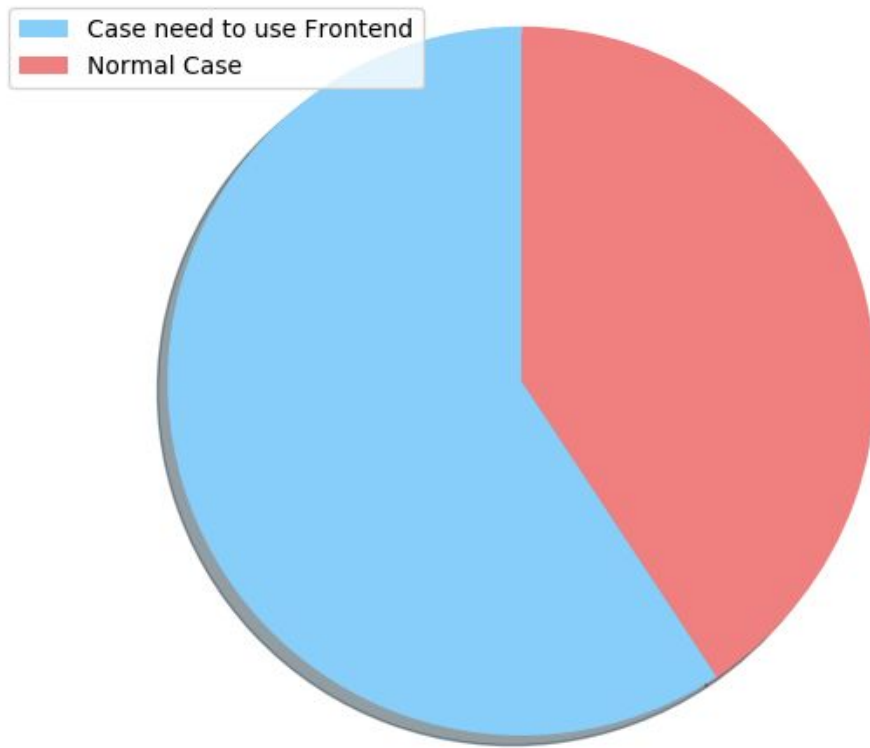
Crawled from

1/2018 – now

Source:

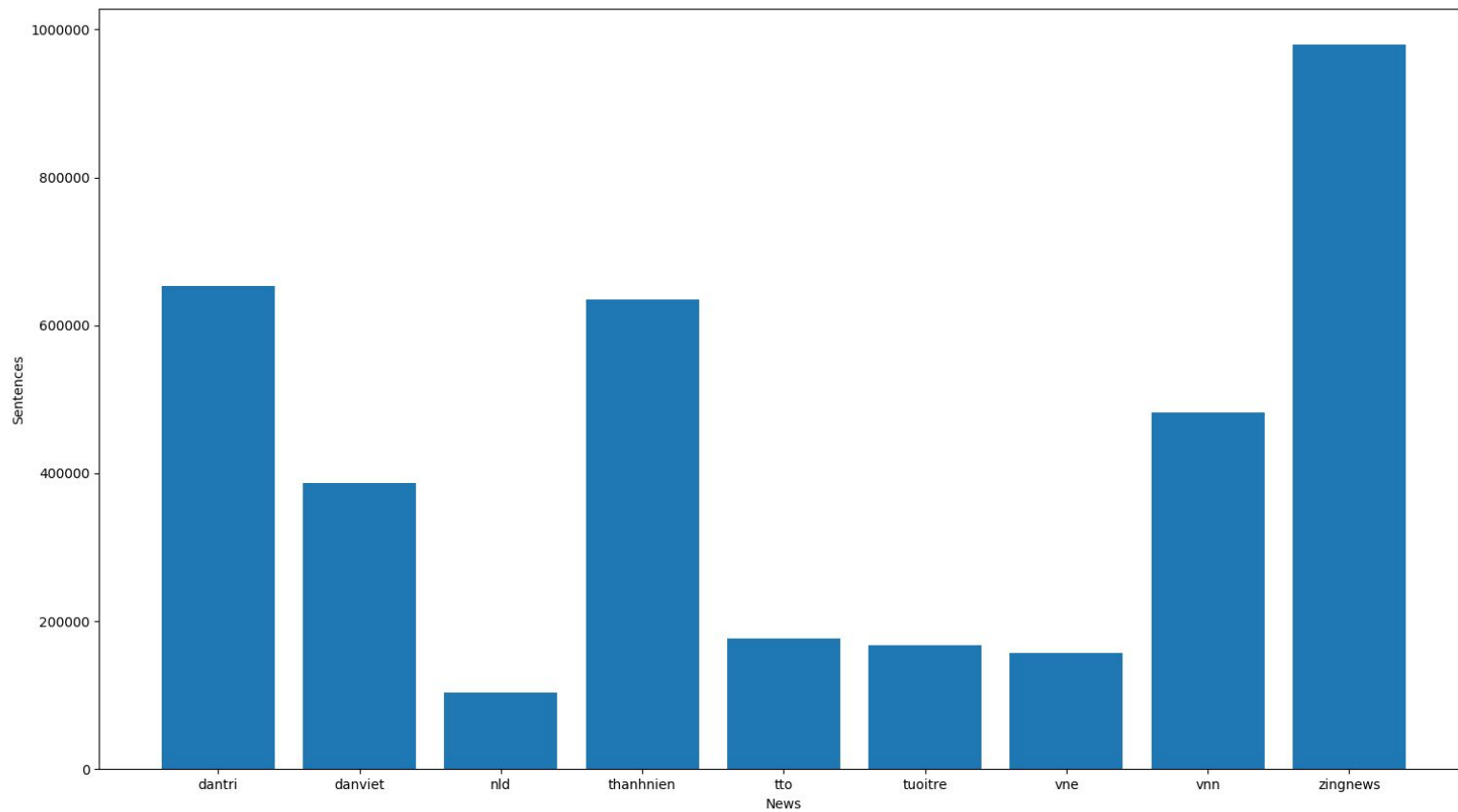
9 online newspapers

News



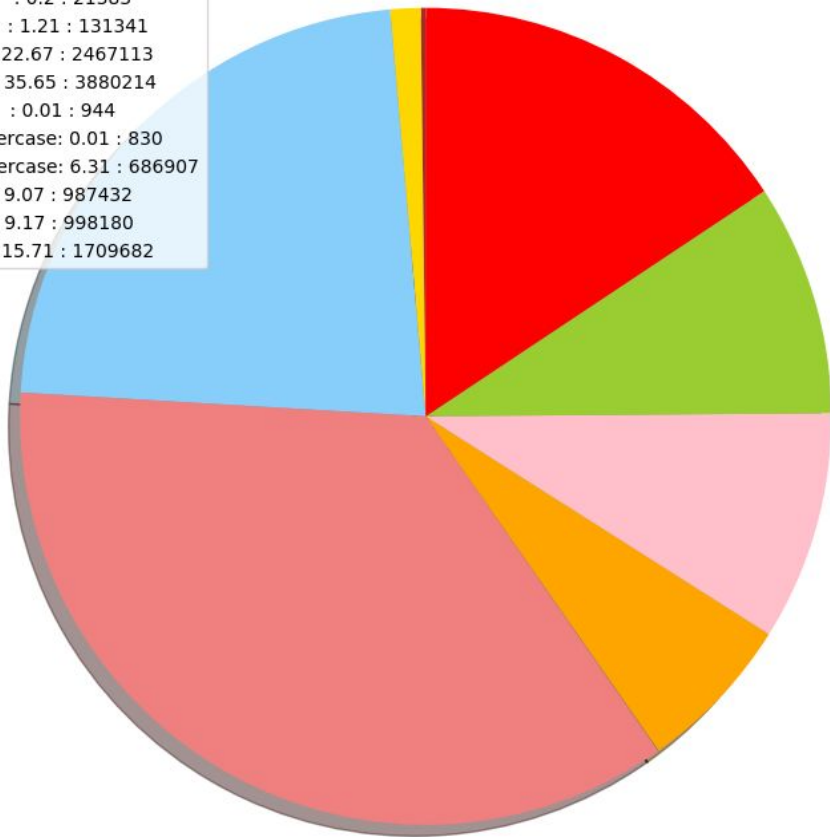
- **Total:** 6.308.173 sentences
- **Not Standardized:** 3.740.507 sentences

Sentences Distribution



[653545, 386444, 103218, 634974, 177287, 167162, 157202, 481566, 979109]

Special Case	: 0.2	: 21383
Timedate	: 1.21	: 131341
Math	: 22.67	: 2467113
English	: 35.65	: 3880214
Teencode	: 0.01	: 944
Acronyms Lowercase	: 0.01	: 830
Acronyms Uppercase	: 6.31	: 686907
Initialism	: 9.07	: 987432
Proper	: 9.17	: 998180
Other	: 15.71	: 1709682



- Special case: website, phone, football, email
- Math: measure, digit, roman
- Initialism: total uppercase but not in English or Acronyms
- Proper: Uppercase first letter, not in dictionary

Insights

- Special Case is just 0.2%
- Most of the Timedate cases are the publication date of the news
- English accounts for 40 percent
- Acronyms Upper and Lower problem

Random Test

60 tricky test cases

To compare TTS systems

100 need-normalized case

Use as baseline to improve

500 random in practical contexts

Random test

Category	Current VOS	Targets	Achieved
Mathematic (6)	50%	66.6%	83.3%
Address (5)	20%	80%	100% (5/5)
Time-Date (15)	73.3%	93.3%	100% (15/15)
Special Case - Measure Unit (15)	66.6%	93.3%	100% (15/15)
Teen Code (5)	20%	100%	80% (4/5)
Acronym (12)	41.6%	100%	100% (12/12)
Upper-lowercase (2)	50%	100%	100%
Cross language	20%		
Proper Noun	20%		

***This is just a survey result on a sample of 60 tests:**

https://docs.google.com/spreadsheets/d/1aXQqJYxlvDwSCUPM4BLiuy3rQUhhdpcSy_Frut-nBIM/edit#gid=0

53.3% -> 96.6%

- Total 60 test case
 - Not include:
 - Normal sentences
 - Cross Language
 - Proper Noun
 - 11 remaining cases:
 - Mathematics
 - Context check => Need statistics from 1GB News
-

Some test cases

- USB wifi hãng **TP** LINK
- số điện thoại **1800.8098**
- trong khoảng thời gian **9h-12h30**
- tương đương 1n \approx 0,1kg)
- Khí **H2S** còn gọi là khí Hidro Sunfua Hóa lỏng ở **60°C**
- Buôn Ma Thuột (**Đăk Lăk**)
- giá từ 500 triệu đến 1 tỷ **đồng/sào**
- cả Tấn Sinh (**1m83**) và Thành Chung (**1m80**) đều là mẫu trung vệ

Test cases from News are being used to improve Rules and expand Dictionary

100 Random case Needs to be standardized

Type	Case	Test (thực tế, bao quát, không trùng lặp, tỉ lệ theo độ phổ biến)	VOS	VOS Update Frontend	Output
20	Mathematic: normal number	Tại cơ quan bảo điện từ Dân trí, sau khi biết tin được bạn đọc giúp đỡ số tiền 285.550.000 đồng			tại cơ quan bảo điện từ dân trí, sau khi biết tin được bạn đọc giúp đỡ số tiền hai trăm tám mươi lăm triệu năm trăm năm mươi nghìn đồng
	Mathematic: normal number	Tổng doanh số bán hàng của toàn thị trường đạt 17.067 xe, trong đó có 11.625 xe du lịch, 4.174 xe thương mại và 180 xe chuyên dụng			tổng doanh số bán hàng của toàn thị đạt mười bảy nghìn không trăm sáu mươi bảy xe, trong đó có mười một nghìn sáu trăm hai mươi lăm xe du lịch, bốn nghìn một trăm bảy mươi tư xe thương mại và một trăm tám mươi xe chuyên dụng
	Mathematic: normal number	quỹ đầu tư vàng lớn thế giới đã bán ra lượng vàng lớn với 21,75 tấn vàng, lượng vàng nắm giữ còn 802,12 tấn.			quỹ đầu tư vàng lớn thế giới đã bán ra lượng vàng lớn với hai mươi một phẩy bảy mươi lăm tấn vàng, lượng vàng nắm giữ còn tám trăm linh hai phẩy mười hai tấn.
	Mathematic: normal number	hậu quả sau: 1- Buộc nộp lại số lợi bất hợp pháp có được do thực hiện hành vi vi phạm hành chính; 2- Buộc thu hồi			hậu quả sau: một buộc nộp lại số lợi bất hợp pháp có được do thực hiện hành vi vi phạm hành chính, hai buộc thu hồi
	Mathematic: normal number + measure	giá vàng giao ngay tại châu Á qua niêm yết có biên độ giảm nhẹ xuống mức 1.313,6 USD/ounce .			giá vàng giao ngay tại châu á qua niêm yết có biên độ giảm nhẹ xuống mức một nghìn ba trăm mười ba phẩy sáu điu ét đi một ao.
RULE	Mathematic: Measure	chính phủ Hàn Quốc ước tính sẽ mất khoảng 3,2 nghìn tỷ won để cung cấp khoảng 2 triệu KW điện cho Triều Tiên			chính phủ hàn quốc ước tính sẽ mất khoảng ba phẩy hai nghìn tỷ quan để cung cấp khoảng hai triệu ki lô quát điện cho triều tiên
	Mathematic: Measure	giá vàng đang được giao dịch ở mức 36,95 triệu đồng/lượng (mua vào) - 37,22 triệu đồng/lượng (bán ra) , tăng tiếp mỗi chiều 70.000 đồng và 170.000 đồng/lượng so với phiên hôm qua.			giá vàng đang được giao dịch ở mức ba mươi sáu phẩy chín mươi lăm triệu đồng, lượng mua vào ba mươi bảy phẩy hai mươi hai triệu đồng, lượng bán ra, tăng tiếp mỗi chiều bảy mươi nghìn đồng và một trăm bảy mươi nghìn đồng, lượng so với phiên hôm qua.
	Mathematic: Measure	Ngôn ngữ nổi lo đầu năm học mới: Phổ biến tình trạng 55 học sinh/lớp			ngôn ngữ nổi lo đầu năm học mới: phổ biến tình trạng năm mươi lăm học sinh, lớp
	Mathematic: Measure	năng suất đất Cần Giờ lên gấp 5-7 lần , cao nhất ngưỡng trên dưới 30 triệu đồng/m²			năng suất đất cần giờ lên gấp năm bảy lần, cao nhất ngưỡng trên dưới ba mươi triệu đồng, m²
	Mathematic: Measure	sở hữu động cơ V8 4.0 lít sản sinh công suất 789 mã lực và 800Nm mô-men xoắn cực đại.			sở hữu động cơ v8 tám béc lít sản sinh công suất bảy trăm tám mươi chín mã lực và tám trăm nơ mô men xoắn cực đại.
	Mathematic: Measure	thu mua 180.000 đồng/kg thay vì gần 670.000 đồng/kg (30 USD/kg)			thu mua một trăm tám mươi nghìn đồng một ki lô gam thay vì gần sáu trăm bảy mươi nghìn đồng một ki lô gam ba mươi điu ét đi, ki lô gam
	Mathematic: Measure	không phát hiện về vấn đề hô hấp đối với Việt, nhưng trái lại em ấy chỉ nặng 48kg và cao 1,60m			không phát hiện về vấn đề hô hấp đối với việt, nhưng trái lại em ấy chỉ nặng bốn mươi tám ki lô gam và cao một phẩy sáu mươi mét
	Mathematic: Measure	Ông Tự được mua với giá rẻ bởi cả ngũ đại dương hiện nay dao động 100.000-120.000 đồng/kg .			ông tự được mua với giá rẻ bởi cả ngũ đại dương hiện nay dao động một trăm nghìn một trăm hai mươi chẵn không không không đồng một ki lô gam.
	Mathematic: phase	tổng sản phẩm quốc gia của Triều Tiên chỉ bằng 1/45 so với Hàn Quốc			tổng sản phẩm quốc gia của triều tiên chỉ bằng một, bốn mươi lăm so với hàn quốc
	Mathematic: phase	Tìm giá trị của biến x để A ≥ k (hoặc A ≤ k, A > k, A < k ...)			tìm giá trị của biến x để a lớn hơn hoặc bằng k hoặc a nhỏ hơn hoặc bằng k, a lớn hơn k, a nhỏ hơn k
	Mathematic: Roman Number	Nói chuẩn bị Đại hội XIII không phải chỉ cho đến năm 2026 mà phải có tầm nhìn chiến lược dài hơn			nói chuẩn bị đại hội mười ba không phải chỉ cho đến năm hai nghìn không trăm hai mươi sáu mà phải có tầm nhìn chiến lược dài hơn

Details: <https://drive.google.com/file/d/1Wg7-hBa0sHluCfPA1BO9ePKzSRU5-o1R/view>

60% -> 97%

- Total 100 test case
 - Not include:
 - Normal sentences
 - Cross Language
 - Proper Noun
 - 11 remaining cases:
 - Wrong mapping acronyms
 - Cannot cover some units
 - Context check => Need statistics from 1GB News
-

500 Random case

CASE	RESULT	OUTPUT
Đặc biệt chú trọng phát triển kinh tế xã hội vùng đồng bào dân tộc thiểu số, miền núi, vùng sâu vùng xa, biên giới, hải đảo thu hẹp khoảng cách phát triển giữa các vùng miền.		đặc biệt chú trọng phát triển kinh tế xã hội vùng đồng bào dân tộc thiểu số, miền núi, vùng sâu vùng xa, biên giới, hải đảo thu hẹp khoảng cách phát triển giữa các vùng miền .
Đan Phong cũng nhiều lần nói về tình cảm đặc biệt anh dành cho con riêng của vợ trên sóng truyền hình.		Đan Phong cũng nhiều lần nói về tình cảm đặc biệt anh dành cho con riêng của vợ trên sóng truyền hình .
Và tôi đây, khi du khách đến trải nghiệm cấp treo Hòn Thơm, sẽ bất ngờ khi thấy một bên căng dầm chất Y – sống động, tươi trẻ và đầy tính nghệ thuật.		và tôi đây, khi du khách đến trải nghiệm cấp treo hòn thơm , sẽ bất ngờ khi thấy một bên căng dầm chất y sống động , tươi trẻ và đầy tính nghệ thuật.
Ngày cả khi Venezuela thực hiện chuyển giao quyền lực thì Trung Quốc chắc chắn sẽ là xuất yếu cầu tại thiết Venezuela, bởi vì họ có nhiều kênh hợp pháp trong nhiều lĩnh vực, đặc biệt là mỏ dầu mỏ cầu orinoco.		ngày cả khi venezuela thực hiện chuyển giao quyền lực thì trung quốc chắc chắn sẽ là xuất yếu cầu tại thiết venezuela , bởi vì họ có nhiều kênh hợp pháp trong nhiều lĩnh vực , đặc biệt là mỏ dầu mỏ cầu orinoco .
Giám đốc Trung tâm Pháp y TP Cần Thơ Hồ Bẩy phát biểu tại buổi làm việc với Ủy ban Tư pháp ngày 16/4.		giám đốc trung tâm pháp y thành phố cần thơ hồ bẩy phát biểu tại buổi làm việc với ủy ban tư pháp ngày mười sáu tháng bốn .
Bất đồng về mức chi tiêu ngân sách và về chính sách nhập cư từng khiến chính phủ của ông Trump đóng cửa 3 ngày hồi tháng 1 năm nay và đóng cửa vài giờ vào tháng 2.		bất đồng về mức chi tiêu ngân sách và về chính sách nhập cư từng khiến chính phủ của ông trump đóng cửa của ba ngày hồi tháng một năm nay và đóng cửa vài giờ vào tháng hai .
Khi có các vụ việc, bức xúc còn mạnh mẽ hơn, tổ chức công đoàn đã phối hợp với giới chủ để cùng giải quyết, tạo mối quan hệ lao động hài hoà” - ông Ngọc Duy hiểu cho biết.		khi có các vụ việc , bức xúc còn mạnh mẽ hơn , tổ chức công đoàn đã phối hợp với giới chủ để cùng giải quyết , tạo mối quan hệ lao động hài hòa ông ngọc duy hiểu cho biết .
Thu nhập bình quân của người lao động 11,401 triệu đồng/người/tháng (đạt 100,9% KHN); trong đó khối sản xuất than: 12,807 triệu đồng/người/tháng, tăng 5,5% so với KHN...		thu nhập bình quân của người lao động , mười một phẩy bốn trăm linh một triệu đồng người xuất than đạt một trăm phẩy chín phần trăm cây éch en , trong đó khối sản xuất than , mười hai phẩy tám trăm linh bảy triệu đồng xuất người xuất than , tăng năm phẩy năm phần trăm so với cây éch en
So với các dự án tương tự trên thế giới, giá của biệt thự HOLM là hợp lý bởi thiết kế chuẩn quốc tế, vị trí đắc địa gần trung tâm và đặc biệt là ưu thế tọa lạc hướng sông với tầm nhìn tuyệt mỹ.		so với các dự án tương tự trên thế giới , giá của biệt thự éch au eo em là hợp lý bởi thiết kế chuẩn quốc tế , vị trí đắc địa gần trung tâm và đặc biệt là ưu thế tọa lạc hướng sông với tầm nhìn tuyệt mỹ .
Trong thời gian Cao Ca sang sống cùng chồng tại Lion, nơi có trụ sở Interpol đã thuê một tòa biệt thự có vườn hoa rộng hàng ngàn mét vuông.		trong thời gian cao ca sang sống cùng chồng tại lion , nơi có trụ sở interpol đã thuê một tòa biệt thự có vườn hoa rộng hàng ngàn mét vuông .
Đề tiếp tục đưa vốn đến gần hơn với người dân khu vực nông thôn, vùng sâu, vùng xa, viết tiếp những thành quả tích cực của agribank đạt được đối với quá trình phát triển kinh tế hộ, ngay trước thềm agribank kỷ niệm 30 năm thành lập, Agribank tích cực triển khai trên toàn hệ thống Điểm giao dịch lưu động bằng xe ô tô chuyên dùng, đáp ứng nhu cầu gửi, vay vốn, cung cấp sản phẩm dịch vụ tiện ích ngân hàng đến khách hàng nhất là vùng sâu vùng xa đi lại khó khăn, được chính quyền, nhân dân các địa phương đánh giá cao, đồng tình ủng hộ và được Ngân hàng Nhà nước đánh giá là nét mới của Agribank trong thực hiện nhiệm vụ gắn với địa bàn vùng sâu vùng xa.		để tiếp tục đưa vốn đến gần hơn với người dân khu vực nông thôn , vùng sâu , vùng xa , viết tiếp những thành quả tích cực của agribank đạt được đối với quá trình phát triển kinh tế hộ , ngay trước thềm agribank kỷ niệm ba mươi năm thành lập , agribank tích cực triển khai trên toàn hệ thống điểm giao dịch lưu động bằng xe ô tô chuyên dùng , đáp ứng nhu cầu gửi , vay vốn , cung cấp sản phẩm dịch vụ tiện ích ngân hàng đến khách hàng nhất là vùng sâu vùng xa đi lại khó khăn , được chính quyền , nhân dân các địa phương đánh giá cao , đồng tình ủng hộ và được ngân hàng nhà nước đánh giá là nét mới của agribank trong thực hiện nhiệm vụ gắn với địa bàn vùng sâu vùng xa .
Làng Chăm Mỹ Nghiệp là một làng nghề dệt thảm truyền thống nổi tiếng ở Ninh Thuận, kẻ bán là làng nghề gốm Bàu Trúc cũng nổi tiếng không kém nên hàng năm thu hút rất đông khách du lịch đến tham quan, nghiên cứu.		làng chăm mỹ nghiệp là một làng nghề dệt thảm truyền thống nổi tiếng ở ninh thuận , kẻ bán là làng nghề gốm bàu trúc cũng nổi tiếng không kém nên hàng năm thu hút rất đông khách du lịch đến tham quan , nghiên cứu .
Emily Blunt nhận giải Nữ phụ xuất sắc với vai diễn trong "A Quiet Place" tại Screen Actors Guild Awards		emily blunt nhận giải nữ phụ xuất sắc với vai diễn trong a quiet place tại screen actors guild awards
Nhân dịp này, Tổng Bí thư, Chủ tịch nước Nguyễn Phú Trọng đã gửi tặng vương miện trăm tuổi cho máy vi tính để quốc vương làm quà tặng cho các cơ sở giáo dục của Campuchia.		nhân dịp này , tổng bí thư , chủ tịch nước nguyên phú trọng đã gửi tặng vương miện trăm tuổi cho máy vi tính để quốc vương làm quà tặng cho các cơ sở giáo dục của campuchia .
Do đó, doanh số của các dòng xe bình dân và trung cấp vào năm 2019 được kỳ vọng sẽ tăng mạnh ở mức 20% so với cùng kỳ.		do đó , doanh số của các dòng xe bình dân và trung cấp vào hai nghìn không trăm mười chín kỳ vọng sẽ mạnh ở mức hai mươi phần trăm so với cùng kỳ .
Phút 54. Đunk đánh đầu đi vọt xa khung thành của Man City từ quả đá phạt của đội chủ nhà, một pha không nguy hiểm với Ederson.		phút năm mươi tư , dunk đánh đầu đi vọt xa khung thành của man city từ quả đá phạt của đội chủ nhà , một pha không nguy hiểm với ederson .
Lâm sao phụ nữ có thể từ chức được thời gian cho việc tập thể dục trong khi vừa phải làm việc tập thể dục ở cơ quan của chăm lo cho con cái ở nhà?		lâm sao phụ nữ có thể từ chức được thời gian cho việc tập thể dục trong khi vừa phải làm việc tập thể dục ở cơ quan của chăm lo cho con cái ở nhà .
Trình tự hiện nay các nhà chung cư cao tầng, các thiết bị của cư dân rất hiện đại nhưng hầu như không đủ chi phí để bảo trì, bảo dưỡng trang thiết bị.		trình tự hiện nay các nhà chung cư cao tầng , các thiết bị của cư dân rất hiện đại nhưng hầu như không đủ chi phí để bảo trì , bảo dưỡng trang thiết bị .
Ông Nguyễn Minh Nguyễn - Phó tổng Giám đốc Công ty cổ phần đầu tư Văn Phú-Invest (thứ hai từ trái sang) đại diện cho đơn vị nhận cơ Đon vị xuất sắc trong phong trào thi đua.		ông nguyên minh nguyên phó tổng giám đốc công ty cổ phần đầu tư văn phú invest thứ hai từ trái sang đại diện cho đơn vị nhận cơ đon vị xuất sắc trong phong trào thi đua .
Ông Kevin Feige đã năm giờ vai trò là chủ tịch của Marvel Studios kể từ năm 2007, ông đã đồng hành cùng với những siêu anh hùng của Marvel ngay từ buổi ban đầu.		ông kevin feige đã năm giờ vai trò là chủ tịch của marvel studios kể từ năm hai nghìn không trăm linh bảy , ông đã đồng hành cùng với những siêu anh hùng của marvel ngay từ buổi ban đầu .
Chia sẻ với Dân trí, Trần Thanh cho biết, bay sang Indonesia làm MC nhưng anh chỉ được bà xã Hari Won "phát lương" vốn 2 triệu đồng.		chia sẻ với dân trí , trần thanh cho biết , bay sang indonesia làm mc nhưng anh chỉ được bà xã hari won "phát lương" vốn hai triệu đồng .
Nổi diễn viên 35 tuổi nổi tiếng sau bộ phim "Vũ điệu hoàng đế" (năm 2007) và từng nhiều lần có tên trong danh sách những mỹ nhân đẹp nhất Philippines của tạp chí FHM hay Maxim.		nổi diễn viên ba mươi lăm tuổi nổi tiếng sau bộ phim vũ điệu hoàng đế năm hai nghìn không trăm linh bảy và từng nhiều lần có tên trong danh sách những mỹ nhân đẹp nhất philippines của tạp chí ép éch em hay maxim .
Từng khuôn viên đều được thiết kế các góc thư giãn, mang lại cho giáo viên và học sinh cảm giác thoải mái sau những giờ học căng thẳng.		từng khuôn viên đều được thiết kế các góc thư giãn , mang lại cho giáo viên và học sinh cảm giác thoải mái sau những giờ học căng thẳng .
Tuần trước, Yoon Ji Oh đã tiết lộ về việc cô luôn cảm thấy bất an khi nhà cô ở gần công trường xây dựng.		tuần trước , yoon ji oh đã tiết lộ về việc cô luôn cảm thấy bất an khi nhà cô ở gần công trường xây dựng .
Mẫu crossover có nhô này của Hyundai đã bị bắt gặp trên đường chạy thử ở nhiều nơi trên thế giới và một số chuyên trang ô tô cho rằng nó có liên quan tới chiếc Carmino concept đã ra mắt tại Ấn Độ cách đây 3 năm.		mẫu crossover có nhô này của hyundai đã bị bắt gặp trên đường chạy thử ở nhiều nơi trên thế giới và một số chuyên trang ô tô cho rằng nó có liên quan tới chiếc carlino concept đã ra mắt tại ấn độ cách đây ba năm .
Tôi hình dung rõ ràng từng chi tiết, từ đám đông người hâm mộ họ reo tên tôi đến các nhà bảo an tôi tránh nhau được từ chuyển với tôi.		tôi hình dung rõ ràng từng chi tiết , từ đám đông người hâm mộ họ reo tên tôi đến các nhà bảo an tôi tránh nhau được từ chuyển với tôi .
Trực tiếp khảo sát và nghiên cứu bức tranh khác trên đã đặc biệt này, giám đốc bảo tàng Adiyaman cho biết bức họa được khắc khá thô sơ nhưng thể hiện rất rõ cảnh con người và động vật thời cổ đại.		trực tiếp khảo sát và nghiên cứu bức tranh khác trên đã đặc biệt này , giám đốc bảo tàng adiyaman cho biết bức họa được khắc khá thô sơ nhưng thể hiện rất rõ cảnh con người và động vật thời cổ đại .
Khi được nói lời nói sau cùng, bị cáo Trần Phương Bình cho rằng bản thân mình vốn là một giáo viên, luôn tâm huyết với ngành Giáo dục, tìm mọi cách để đưa ngân hàng đi lên.		khi được nói lời nói sau cùng , bị cáo trần phương bình cho rằng bản thân mình vốn là một giáo viên , luôn tâm huyết với ngành giáo dục , tìm mọi cách để đưa ngân hàng đi lên .
Về vụ việc này, cơ quan CSĐT đang xác minh làm rõ việc chuyển tiền giữa bị can Phạm Thanh Liêm và ông D. là giao dịch bình thường, hợp pháp hay có khuất tất gì không?		về vụ việc này , cơ quan cảnh sát điều tra đang xác minh làm rõ việc chuyển tiền giữa bị can phạm thanh liêm và ông d . là giao dịch bình thường , hợp pháp hay có khuất tất gì không ?

480/500

Details: https://drive.google.com/file/d/1x1Uhu8oNaCj-Ql8SXoNt3td8VysWG_q/view

Failed Test Cases

<p>Dền Saturn được người La Mã xây dựng vào năm 497 TGN dưới thời vua Tarquinius Superbus với mục đích thờ thần Saturn.</p> <p>Cũng theo kết quả khảo sát cho thấy, lĩnh vực nghiên cứu của các NNC cũng không đồng đều và có sự chênh lệch khá lớn tập trung nhiều nhất vào 3 lĩnh vực khoa học tự nhiên, khoa học kỹ thuật và công nghệ, khoa học xã hội và nhân văn.</p> <p>Nhận lời mời của Tổng Bí thư, Chủ tịch nước Nguyễn Phú Trọng, Chủ tịch Đảng Lao động Triều Tiên, Chủ tịch Ủy ban Quốc vụ nước Cộng hòa dân chủ nhân dân Triều Tiên Kim Jong-un có chuyến thăm hữu nghị chính thức Việt Nam từ ngày 1/3 - 2/3.</p> <p>Cũng theo người đứng đầu BĐ-TB&XH, sau 3 năm kiến trì, đặc biệt năm 2017 với sự chỉ đạo quyết liệt của Thủ tướng Chính phủ, Bộ đã làm việc với phía Hàn Quốc trên tinh thần quyết liệt xử lý các doanh nghiệp phía bạn vi phạm.</p> <p>Như vậy, tính theo thị giá thì 5,2 triệu cổ phiếu GTT của bà Võ Thị Thanh chỉ có giá trị hơn 2 tỷ đồng.</p> <p>Manhattan Địa chỉ: Regent Hotel - 1 Cuscaden Road Level 2 Giờ mở cửa: Thứ 2-7 từ 17:00 - 1:00; Chủ nhật từ 12:00 - 15:00 và 17:00 - 1:00.</p> <p>Sự kiện thu hút các chuyên gia hàng đầu về trí tuệ nhân tạo (AI), tự động hóa, IoT người Việt đang làm việc trong các tổ chức, doanh nghiệp lớn tại nước ngoài.</p> <p>Về vụ việc này, Cơ quan CSĐT đang xác minh làm rõ việc chuyển tiền giữa bị can Phạm Thanh Liêm và ông D. là giao dịch bình thường, hợp pháp hay có khuất tất gì không?</p> <p>Chủ trương của Bộ GD&ĐT sắp tới sẽ làm rất mạnh về kiểm định chất lượng cơ sở giáo dục và sẽ có thêm những tổ chức kiểm định độc lập (theo Luật GD ĐH sửa đổi).</p> <p>Toyota Hilux 2.8 G 4x4 AT MLM có nội thất da với ghế lái điều khiển điện, hệ thống điều hòa tự động với cửa gió cho hàng ghế sau, hệ thống giải trí với màn hình cảm ứng 7" có hỗ trợ đầu đọc DVD và camera lùi đi kèm các hỗ trợ kết nối USB, bluetooth, AUX... Ngoài hệ thống điều khiển hành trình, mẫu xe này còn có thêm hệ thống khởi động bằng nút bấm, cửa kính chống kẹt (toàn bộ cửa trên xe)...</p> <p>Trong khi Pixel 3a XL được trang bị thời pin có dung lượng 3.700mAh thì phiên bản Pixel 3a chỉ có thời pin 3.000mAh.</p> <p>Vải MV+ thuộc thương hiệu American Tourister, với hơn 80 năm kinh nghiệm trong việc cho ra những sản phẩm phách cách, không ngừng sáng tạo, hợp thời trang và đậm chất Mỹ.</p> <p>Được biết, Đề án "Đầu tư xây dựng các thiết chế của công đoàn tại các khu công nghiệp, khu chế xuất" được Thủ tướng Chính phủ Nguyễn Xuân Phúc ban hành Quyết định số 655/QĐ-TTg vào ngày 12/5/2017, với mục tiêu tổng quát là giao cho Tổng Liên đoàn Lao động Việt Nam đầu tư xây dựng nhà ở, nhà trẻ, siêu thị và các công trình văn hóa, thể thao tại các khu công nghiệp, khu chế xuất.</p> <p>Năm bắt được xu hướng này, quý III-2018 vừa qua đã có 2 hãng hàng không nội địa công bố ứng dụng phương thức thanh toán QR Pay, nhằm đáp ứng các nhu cầu thanh toán ngày càng cao của khách hàng trong thời đại 4.0.</p> <p>(33 tuổi, ngụ xã Vĩnh Mỹ B, huyện Hòa Bình) đang đi xe máy trên tuyến đường thuộc xã Minh Diệu (huyện Hòa Bình) thì bất ngờ bị sét đánh trúng, khiến anh Kh.</p> <p>Theo thông tin ban đầu, vào lúc 16h20 ngày 25/2, tàu BD 94005 TS, công suất 330 CV, trên tàu có 15 người, do ông Phan Mỹ Na (ở Mỹ Thành, huyện Phú Mỹ, Bình Định) làm thuyền trưởng.</p> <p>Đại diện chủ đầu tư cho biết, vì chất lượng khu văn phòng đảm bảo tiêu chuẩn quốc tế nên trong thời gian tới, trụ sở của Tập đoàn Sunshine Group sẽ được chuyển từ tòa nhà Keangnam Landmark72 về Sunshine Center.</p>	<p>dền saturn được người la mã xây dựng vào năm bốn trăm chín mươi bảy tiêu chuẩn ngành dưới thời vua tarquinius superbus với mục đích thờ thần saturn .</p> <p>cũng theo kết quả khảo sát cho thấy, lĩnh vực nghiên cứu của các nhà nguyên căn cũng không đồng đều và có sự chênh lệch khá lớn tập trung nhiều nhất vào ba lĩnh vực khoa học tự nhiên , khoa học kỹ thuật và công nghệ , khoa học xã hội và nhân văn ,</p> <p>nhận lời mời của tổng bí thư , chủ tịch nước nguyên phi trọng , chủ tịch đảng lao động triều tiên , chủ tịch ủy ban quốc vụ nước cộng hòa dân chủ nhân dân triều tiên kim jong un có chuyến thăm hữu nghị chính thức việt nam từ ngày một tháng ba hai , ba .</p> <p>cũng theo người đứng đầu bộ lao động thông báo và xã hội , sau ba năm kiến trì , đặc biệt năm hai nghìn không trăm mười bảy với sự chỉ đạo quyết liệt của thủ tướng chính phủ , bộ đã làm việc với phía hàn quốc trên tinh thần quyết liệt xử lý các doanh nghiệp phía bạn vi phạm .</p> <p>như vậy , tính theo thị giá thì năm phẩy hai triệu cổ phiếu giấy thể thao của bà võ thị thanh chỉ có giá trị hơn hai tỷ đồng .</p> <p>manhattan địa chỉ . regent hotel một cuscaden road level hai giờ mở cửa . thứ hai bảy từ mười bảy một , chủ nhật từ mười hai mười lăm và mười bảy một .</p> <p>sự kiện thu hút các chuyên gia hàng đầu về trí tuệ nhân tạo ai , tự động hóa , iot người việt đang làm việc trong các tổ chức , doanh nghiệp lớn tại nước ngoài .</p> <p>về vụ việc này , cơ quan cảnh sát điều tra đang xác minh làm rõ việc chuyển tiền giữa bị can phạm thanh liêm và ông năm trăm . là giao dịch bình thường , hợp pháp hay có khuất tất gì không .</p> <p>chủ trương của bộ giáo dục và đào tạo sắp tới sẽ làm rất mạnh về kiểm định chất lượng cơ sở giáo dục và sẽ có thêm những tổ chức kiểm định độc lập theo luật giáo dục đại học sửa đổi .</p> <p>toyota hilux hai . tám giờ bốn ịch bốn ày ti em eo em có nội thất da với ghế lái điều khiển điện , hệ thống điều hòa tự động với cửa gió cho hàng ghế sau , hệ thống giải trí với màn hình cảm ứng bảy có hỗ trợ đầu đọc di vi di và camera lùi đi kèm các hỗ trợ kết nối diu ét bi , bluetooth , ày diu it ngoài hệ thống điều khiển hành trình , mẫu xe này còn có thêm hệ thống khởi động bằng nút bấm , cửa kính chống kẹt toàn bộ cửa trên xe</p> <p>trong khi pixel ba bốn mươi được trang bị thời pin có dung lượng ba . bảy trăm mớ a hắt thì phiên bản pixel ba a chỉ có thời pin ba . không không không mớ a hắt .</p> <p>vali một nghìn không trăm linh năm cộng thuộc thương hiệu american tourister , với hơn tám mươi năm kinh nghiệm trong việc cho ra những sản phẩm phách cách , không ngừng sáng tạo , hợp thời trang và đậm chất mỹ .</p> <p>được biết , đề án đầu tư xây dựng các thiết chế của công đoàn tại các khu công nghiệp , khu chế xuất được thủ tướng chính phủ nguyên xuân phúc ban hành quyết định số sáu trăm năm mươi lăm , ki u đề tăng thiết giáp vào ngày mười hai tháng năm năm hai nghìn không trăm mười bảy , với mục tiêu tổng quát là giao cho tổng liên đoàn lao động việt nam đầu tư xây dựng nhà ở , nhà trẻ , siêu thị và các công trình văn hóa , thể thao tại các khu công nghiệp , khu chế xuất .</p> <p>năm bắt được xu hướng này , quý i i i , hai nghìn không trăm mười tám vừa qua đã có hai hãng hàng không nội địa công bố ứng dụng phương thức thanh toán ki a pay , nhằm đáp ứng các nhu cầu thanh toán ngày càng cao của khách hàng trong thời đại bốn . không .</p> <p>ba mươi ba tuổi , ngụ xã vĩnh mỹ b , huyện hòa bình đang đi xe máy trên tuyến đường thuộc xã minh diệu huyện hòa bình thì bất ngờ bị sét đánh trúng , khiến anh khoa học .</p> <p>theo thông tin ban đầu , vào lúc mười sáu giờ hai mươi ngày hai mươi lăm tháng hai , tàu bui định chín mươi tư nghìn không trăm linh năm tiên sĩ , công suất ba trăm ba mươi một trăm linh năm , trên tàu có mười lăm người , do ông phan my na ở mỹ thành , huyện phú mỹ , bình định làm thuyền trưởng .</p> <p>đại diện chủ đầu tư cho biết , vì chất lượng khu văn phòng đảm bảo tiêu chuẩn quốc tế nên trong thời gian tới , trụ sở của tập đoàn sunshine group sẽ được chuyển từ tòa nhà keangnam là ở nò đề mớ a rờ ca bảy mươi hai về sunshine center .</p>
--	---

Main reason

- Wrong mapping acronyms
- Wrong Roman Numbers

After fix: 497/500 remain 3 case

- tàu **BĐ 94005 TS**
- tòa nhà Keangnam **Landmark72**
- Toyota Hilux **2.8 G** 4x4 AT MLM

Details: <https://drive.google.com/file/d/1ntWycuAOF5728-tRSor8NCcrL39CfxNO/view>

Future Work

Context for Ambiguous Cases

- Context for each Rule
- Context for Unit
 - Nhân dân tệ - Yên
- Context for Acronyms
- Context for upper-lowercase

Time date

Classification

- 10/2019
- 10.2019
- 10-2019
- 10,2019

Address-Code

Mathematics

Acronyms

There should be two kind of acronyms in this dictionary:

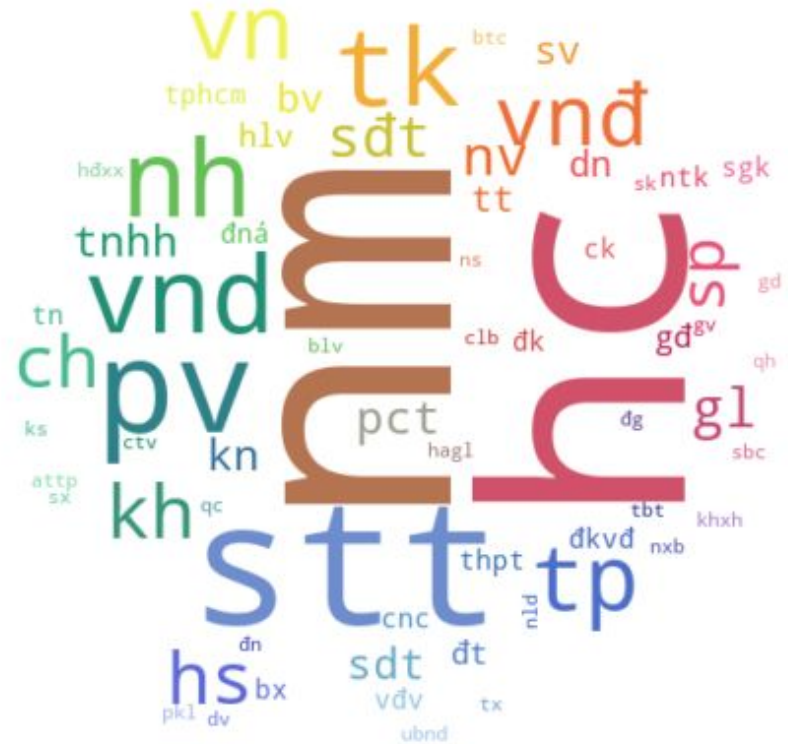
- Need to replace with full meaning phrase:
 - blv -> Bình luận viên ; btc -> ban tổ chức/bộ tài chính/bitcoin
- Need to keep to spell each letter:
 - MV, *WHO, WTO -> vê kép tê ô
- We know exactly they are acronyms, to distinct with unidentified words

Some Acronyms in VOS:

- pq, qx, hc, gl, nm, d̄k, qv

Improve:

- Acronyms Dict much contain Upper and Lower
- An Acronyms may have more than one meaning (have different context):
 - STT, BTC, ₪K...
- Update context for lowercase acronyms



66 Acronyms in lowercase from News 1Gb

- File news 12GB from on Github

- <https://github.com/binhvq/news-corpus>

- 14.896.998 articles

- Mapping for English letter problem:

- O, J, G

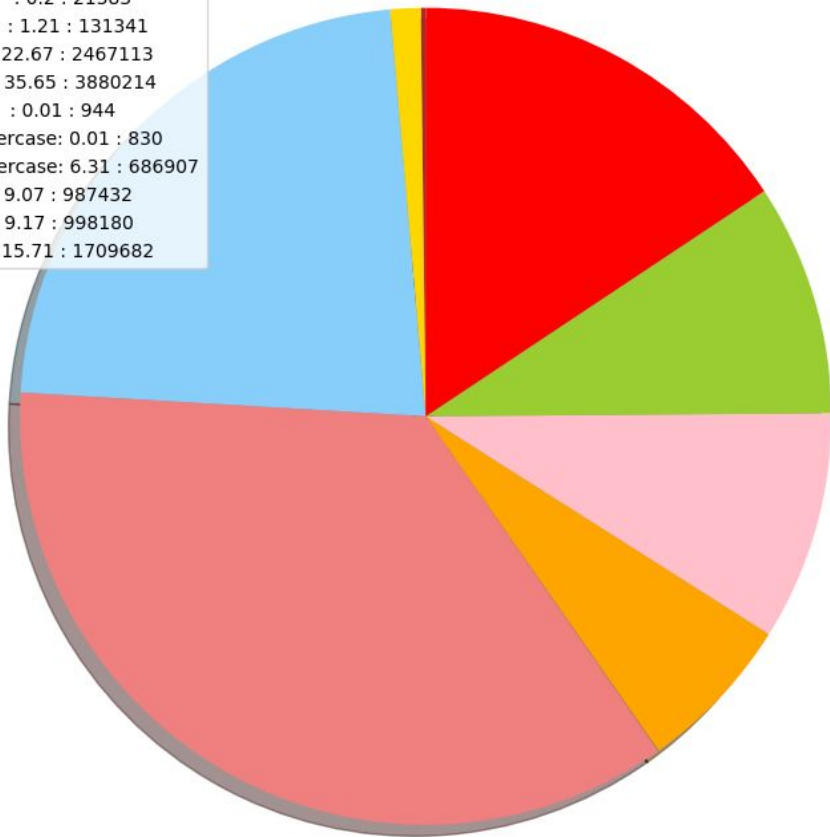
- There is no Vietnamese transliteration

- When to spell Vietnamese when to spell in English

English

English accounts for 40% of sentences need to be normalized.

Special Case	: 0.2	: 21383
Timedate	: 1.21	: 131341
Math	: 22.67	: 2467113
English	: 35.65	: 3880214
Teencode	: 0.01	: 944
Acronyms Lowercase	: 0.01	: 830
Acronyms Uppercase	: 6.31	: 686907
Initialism	: 9.07	: 987432
Proper	: 9.17	: 998180
Other	: 15.71	: 1709682



- Special case: website, phone, football, email
- Math: measure, digit, roman
- Initialism: total uppercase but not in English or Acronyms
- Proper: Uppercase first letter, not in dictionary

Insights

- Special Case is just 0.2%
- Most of the Timedate cases are the publication date of the news
- English accounts for 40 percent
- Acronyms Upper and Lower problem

How Amazon Polly works with these case:

- Particular words, such as company names,
- Acronyms
- Foreign words
- Neologisms (e.g., “ROTFL”, “C’est la vie”)

Use custom lexicons to customize the pronunciation of **Nguyen**

```
<lexeme>
  <grapheme>Nguyen</grapheme>
  <grapheme>nguyen</grapheme>
  <grapheme>NGUYEN</grapheme>
  <phoneme>"nu.jEn"</phoneme>
</lexeme>
```

Phonetic pronunciation for specific text

International Phonetic Alphabet (IPA)

```
<speak>
  You say, <phoneme alphabet="ipa" ph="pɪ'kɑ:n">pecan</phoneme>.
  I say, <phoneme alphabet="ipa" ph="'pi.kæn">pecan</phoneme>.
</speak>
```

Extended Speech Assessment Methods Phonetic Alphabet (X-SAMPA)

```
<speak>
  You say, <phoneme alphabet='x-sampa' ph='pI"kA:n'>pecan</phoneme>.
  I say, <phoneme alphabet='x-sampa' ph='""pi.k{n'>pecan</phoneme>.
</speak>
```

Cross Language (English)

- Popular English:
 - Singapore->sin ga bo
- With other words: (have meaning and available in Language Dictionary)



This is not totally standards

Understand -> ən.dʒ:.'stænd

```
INSERT INTO `ipa` (`ID`, `word`, `ipa`) VALUES
(1, 'ABC', '[eibi:si:]'),
(2, 'ABC book', '[eibi:si:buk]'),
(3, 'AIDS', '[eidz]'),
(4, 'Abkhazia', '[əbkheiziə]'),
(5, 'Abkhazian', '[əbkheizin]'),
(6, 'Abyssinia', '[əbisiniə]'),
(7, 'Abyssinian', '[əbisiniən]'),
(8, 'Acheron', '[ətferən]'),
(9, 'Achilles', '[ətʃilz]'),
(10, 'Achilles' heel', '[ətʃilshi:l]'),
(11, 'Achilles' tendon', '[ətʃilstendən]'),
(12, 'Adam', '[ædəm]'),
(13, 'Addis Ababa', '[ədaizæbəbə]'),
(14, 'Aden', '[ədn]'),
(15, 'Adonis', '[ədɒni:z]'),
(16, 'Adrianople', '[ədraiənəpl]'),
(17, 'Adriatic', '[ədriætik]'),
(18, 'Adriatic Sea', '[ədriætiksia]'),
(19, 'Adzharia', '[ædzɛəriə]'),
(20, 'Aegean', '[i:dʒiən]'),
```

8752 words

Link download IPA <https://sourceforge.net/projects/free-english-to-ipa-database/>

App can speech IPA <https://play.google.com/store/apps/details?id=com.hoardingsinc.pronunroid>



Pronunroid - IPA pronunciation

Hoardings Inc. Education

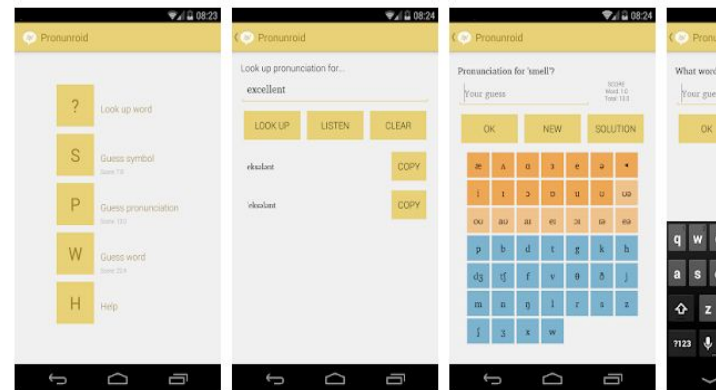
★★★★★ 1,098

3+

This app is compatible with your device.

Add to Wishlist

Install



Sticky Phrase

<https://github.com/coccoc/coccoc-tokenizer>

- In website:
 - thegioididong.com
- In email:
 - nguyenvana@gmail.com

Some note

Paper link: <https://www.overleaf.com/project/5f2114e601f3f40001f7a805>

Outline:

+Introduction:

+Related Work:

- Giới thiệu qua lịch sử TTS

- Giới thiệu Tacotron2 model: Các nét đặc trưng

+Method: Ứng dụng cho tiếng Việt

- Vinorm

- Viphoneme

+Thống kê và kết quả

- 2 phần thống kê trong paper

- Thêm cái MCD: on 200 để đảm bảo chính xác hôm bị review: 691,9

- Demo: <http://mostts.herokuapp.com/>

+Future Work