

Yepoko Lessons For Machine Learning on Small Data



Xuanyi Chew
YOW! Data 2021

The Puzzle

Term	Number
rureponga talu	10
malapunga yepoko	15
supu	20
tokapunga telu	21
alapunga yepoko	27
polangipula talu	30
tokapu rureponga yepoko	35
tokapu malapu	40
tokapu talu	48
tokapu alapunga talu	50
tokapu talu tokapunga telu	69
tokapu talu polangipunga yepoko	79
tokapu yepoko alapunga telu	97

The Puzzle

Term	Number
rureponga talu	10
malapunga yepoko	15
supu	20
tokapunga telu	21
alapunga yepoko	27
polangipula talu	30
tokapu rureponga yepoko	35
tokapu malapu	40
tokapu talu	48
tokapu alapunga talu	50
tokapu talu tokapunga telu	69
tokapu talu polangipunga yepoko	79
tokapu yepoko alapunga telu	97

Translate:

- tokapu polangipu
- tokapu talu rureponga telu
- tokapu yepoko malapunga talu
- tokapu yepoko polangipunga telu
- 13
- 66
- 72
- 76
- 95

The Puzzle - One More Thing

Term	Number
rureponga talu	10
malapunga yepoko	15
supu	20
tokapunga telu	21
alapunga yepoko	27
polangipula talu	30
tokapu rureponga yepoko	35
tokapu malapu	40
tokapu talu	48
tokapu alapunga talu	50
tokapu talu tokapunga telu	69
tokapu talu polangipunga yepoko	79
tokapu yepoko alapunga telu	97

Translate:

- tokapu polangipu
- tokapu talu rureponga telu
- tokapu yepoko malapunga talu
- tokapu yepoko polangipunga telu
- 13
- 66
- 72
- 76
- 95

telu < yepoko

Neural Language Models to the Rescue

The ML Tasks

1. Categorization
2. Regression
3. Translation (sequence to sequence)

Neural Language Models: The Process

1. Download a pre-trained model
2. Fine-tune it to the task using examples (all 13 of them!)
3. ???
4. Profit!!!

BERT Results

Test	Predicted	Result
tokapu polangipu	55.21	X
tokapu talu rureponga telu	54.44	X
tokapu yepoko malapunga talu	57.98	X
tokapu yepoko polangipunga telu	58.22	X

GPT-3

> “If two times three is six, what is four times four”

“Sixteen”

GPT-3

> if 'rureponga talu' means 10 and 'malapunga yepoko' means 15 and 'supu' means 20 and 'tokapunga telu' means 21 and 'alapunga yepoko' means 27 and 'polangipula talu' means 30 and 'tokapu rureponga yepoko' means 35 and 'tokapu malapu' means 40 and 'tokapu talu' means 48 and 'tokapu alapunga talu' means 50 and 'tokapu talu tokapunga telu' means 69 and 'tokapu talu polangipunga yepoko' means 79 and 'tokapu yepoko alapunga telu' means 97, then what does 'tokapu polangipu' mean?

"I understand! It means 'twice 10!'!"

> “what does ‘tokapu talu rureponga telu’ mean?”

"Ah! I know the answer to this! It's 'uproot 100 10 rods long trees!'!"

GPT-3

> if 'rureponga talu' means 10 and 'malapunga yepoko' means 15 and 'supu' means 20 and 'tokapunga telu' means 21 and 'alapunga yepoko' means 27 and 'polangipula talu' means 30 and 'tokapu rureponga yepoko' means 35 and 'tokapu malapu' means 40 and 'tokapu talu' means 48 and 'tokapu alapunga talu' means 50 and 'tokapu talu tokapunga telu' means 69 and 'tokapu talu polangipunga yepoko' means 79 and 'tokapu yepoko alapunga telu' means 97, then what does 'tokapu talu rureponga telu' mean?

"I know the answer! Here's a guess: 100 rods long, 10 inches in diameter, made of wood, with three holes."

> "what does 'tokapu polangipu' mean?"

"I think it means 'two times 100'."

GPT-3

> if 'rureponga talu' means 10 and 'malapunga yepoko' means 15 and 'supu' means 20 and 'tokapunga telu' means 21 and 'alapunga yepoko' means 27 and 'polangipula talu' means 30 and 'tokapu rureponga yepoko' means 35 and 'tokapu malapu' means 40 and 'tokapu talu' means 48 and 'tokapu alapunga talu' means 50 and 'tokapu talu tokapunga telu' means 69 and 'tokapu talu polangipunga yepoko' means 79 and 'tokapu yepoko alapunga telu' means 97, then what does 'tokapu polangipu' mean?

Let me take a look at my old records. Yes, 'tokapu polangipu' means 100.

> what does 'tokapu talu rureponga telu' mean?

Sorry, I can't find the answer for this.

GPT-3

...

Socrates: tokapu alapunga talu

Aristotle: 50

Socrates: tokapu talu tokapunga telu

Aristotle: 69

Socrates: tokapu talu polangipunga yepoko

Aristotle: 79

Socrates: tokapu yepoko alapunga telu

Aristotle: 97

Socrates: tokapu polangipu

Aristotle:

I think it means 'Socrates is a dung beetle, Aristotle is a snake, and I am a dodo.'

Random Forests

1. Byte-pair encode text to be used as features
2. Encode position of pairs in sentence
3. ???
4. Profit!

Random Forests Feature Matrix

.	C:Syl1	N:Pos1	C:Syl2	N:Pos2	C:Syl3	N:Pos3	...	C:Syl12	N:Pos12	N:Answer
1	146	0	145	0	129	0	...	0	0	10
2	144	0	136	0	121	1	...	0	0	15
3	115	0	117	0	140	0	...	0	0	20
4	148	0	135	0	136	0	...	0	0	21
5	142	0	136	0	121	1	...	0	0	27
6	150	0	149	0	137	0	...	0	0	30
7	148	0	135	0	146	1	...	0	0	35
8	148	0	135	0	144	1	...	0	0	40
9	148	0	135	0	116	1	...	0	0	48
10	148	0	135	0	142	1	...	0	0	50
11	148	0	135	0	116	1	...	0	0	69
12	148	0	135	0	116	1	...	111	3	79
13	148	0	135	0	121	1	...	0	0	97

Random Forests - Results

Test	Predicted	Result
tokapu polangipu	36.74	X
tokapu talu rureponga telu	61.08	X
tokapu yepoko malapunga talu	57.62	X
tokapu yepoko polangipunga telu	52.09	X

Human (me)

- Finished translation in about an 40 mins.
- Used basic statistics.
- Required basic linguistics knowledge.
- Required backtracking.
- Required pattern matching.
- Required basic arithmetics.
- Required basic algebra.

How I Solved It

Offshoot On Orthography

Term	Number
ၚုၚေၚံၚာတၚ်အိၚ်	10
ၚာအိၚ်အၚ်ၚာၚေၚံၚာ	15
တၚ်ဟံၚ	20
တၚ်ၚာၚေၚံၚာတၚ်အိၚ်	21
အိၚ်အၚ်ၚာၚေၚံၚာ	27
ၚာၚ်အိၚ်အၚ်ၚာၚေၚံၚာတၚ်အိၚ်	30
တၚ်ၚာၚေၚံၚာၚုၚေၚံၚာၚေၚံၚာ	35
တၚ်ၚာၚေၚံၚာၚာအိၚ်အၚ်	40
တၚ်ၚာၚေၚံၚာတၚ်အိၚ်	48
တၚ်ၚာၚေၚံၚာအိၚ်အၚ်ၚာၚေၚံၚာတၚ်အိၚ်	50
တၚ်ၚာၚေၚံၚာတၚ်အိၚ်တၚ်ၚာၚေၚံၚာတၚ်အိၚ်	69
တၚ်ၚာၚေၚံၚာတၚ်အိၚ်ၚာၚ်အိၚ်အၚ်ၚာၚေၚံၚာ	79
တၚ်ၚာၚေၚံၚာၚေၚံၚာအိၚ်အၚ်ၚာၚေၚံၚာတၚ်အိၚ်	97



Term	Number
rureponga talu	10
malapunga yepoko	15
supu	20
tokapunga telu	21
alapunga yepoko	27
polangipula talu	30
tokapu rureponga yepoko	35
tokapu malapu	40
tokapu talu	48
tokapu alapunga talu	50
tokapu talu tokapunga telu	69
tokapu talu polangipunga yepoko	79
tokapu yepoko alapunga telu	97

Syllables

alapunga

Syllables

alapunga

[a, kɿ̥a, βʊ, ŋa]

Byte Pair Encoding

alapunga

Byte Pair Encoding

alapunga

[aɪ, ɪa, ap, pu, un, ng, ga]

Byte Pair Encoding

alapunga

[aɫ, ɫa, ap, pu, un, ng, ga]

[🇷🇺, 💎, 🙌]

Byte Pair Encoding

alapunga

[aɫ, ɫa, ap, pu, un, ng, ga]

[🇷🇺, 💎, 🙌]

🇷🇺	aɫ
💎	Z u
🙌	NG a
Z	ap
NG	ng

Byte Pair Statistics

pu	17	la	8	gi	2
ap	14	ep	7	ip	2
ok	14	un	7	ma	2
al	11	ta	6	ol	2
ng	11	ko	5	on	2
ga	9	ye	5	re	2
ka	9	el	3	ru	2
lu	9	te	3	su	1
po	9	ur	2	ul	1
to	9	an	2	up	1

Syllable Statistics

pu	17	la	8	gi	2		
ap	14	ep	7	ip	2		
ok	14	un	7	ma	2		
al	11	ta	6	ol	2		
ng	11	ko	5	on	2	nga	9
ga	9	ye	5	re	2	ngi	2
ka	9	el	3	ru	2	a	2
lu	9	te	3	su	1		
po	9	ur	2	ul	1		
to	9	an	2	up	1		

Syllable Statistics

pu	17	la	8	gi	2		
ap	14	ep	7	ip	2		
ok	14	un	7	ma	2		
al	11	ta	6	ol	2		
ng	11	ko	5	on	2	nga	9
ga	9	ye	5	re	2	ngi	2
ka	9	el	3	ru	2	a	2
lu	9	te	3	su	1		
po	9	ur	2	ul	1		
to	9	an	2	up	1		

Recursive Pattern Matching

Term	Number
rureponga talu	10
malapunga yepoko	15
supu	20
tokapunga telu	21
alapunga yepoko	27
polangipula talu	30
tokapu rureponga yepoko	35
tokapu malapu	40
tokapu talu	48
tokapu alapunga talu	50
tokapu talu tokapunga telu	69
tokapu talu polangipunga yepoko	79
tokapu yepoko alapunga telu	97

Some “phrases” are repeated.

Recursive Pattern Matching

Term	Number
rureponga talu	10
malapunga yepoko	15
supu	20
tokapunga telu	21
alapunga yepoko	27
polangipula talu	30
tokapu rureponga yepoko	35
tokapu malapu	40
tokapu talu	48
tokapu alapunga talu	50
tokapu talu tokapunga telu	69
tokapu talu polangipunga yepoko	79
tokapu yepoko alapunga telu	97

Some “phrases” are repeated.

We now have “word” units.

Recursive Pattern Matching

Term	Number
rureponga talu	10
malapunga yepoko	15
supu	20
tokapunga telu	21
alapunga yepoko	27
polangipula talu	30
tokapu rureponga yepoko	35
tokapu malapu	40
tokapu talu	48
tokapu alapunga talu	50
tokapu talu tokapunga telu	69
tokapu talu polangipunga yepoko	79
tokapu yepoko alapunga telu	97

Some “phrases” are repeated.

We now have “word” units.

We now have “sub-word” units.

Solving It - Apply Broad Pattern Matching

tokapunga telu	21
tokapu talu	48
tokapu talu tokapunga telu	69

+ Bigrams should be considered.

tokapu talu tokapunga telu

48

21

69

* - assumption
? - open question
+ - newly synthesized fact

Solving It - First Level Pattern Matching

tokapunga telu	21
tokapu talu	48
tokapu talu tokapunga telu	69

+ Bigrams should be considered.

+ Juxtaposition of bigrams implies addition.

? How tightly do bigrams bind?

tokapu talu tokapunga telu

48 + 21

69

* - assumption
? - open question
+ - newly synthesized fact

Solving It - Apply New Information on New Problem

tokapunga telu	21
tokapu talu	48
tokapu talu tokapunga telu	69
tokapu talu polangipunga yepoko	79

+ Bigrams should be considered.

+ Juxtaposition of bigrams implies addition.

? How tightly do bigrams bind?

tokapu talu polangipunga yepoko

48 + x

79

* - assumption
? - open question
+ - newly synthesized fact

Solving It - Apply New Information on New Problem

tokapunga telu	21
tokapu talu	48
tokapu talu tokapunga telu	69
tokapu talu polangipunga yepoko	79

+ Bigrams should be considered.

+ Juxtaposition of bigrams implies addition.

? How tightly do bigrams bind?

tokapu talu polangipunga yepoko

48 + 31

79

polangipunga yepoko	31
---------------------	----

* - assumption
? - open question
+ - newly synthesized fact

Solving It - A Leap of Faith

polangipunga yepoko	31
polangipula talu	30

+ Bigrams should be considered.

+ Juxtaposition of bigrams implies addition.

? How tightly do bigrams bind?

* polangipula is a typo of polangipunga

polangipunga yepoko	31
---------------------	----

* - assumption
? - open question
+ - newly synthesized fact

Solving It - A Second Leap of Faith

polangipunga yepoko	31
polangipula talu	30

polangipunga yepoko = 31
polangipunga talu = 30

+ Bigrams should be considered.

+ Juxtaposition of bigrams implies addition.

? How tightly do bigrams bind?

* polangipula is a typo of polangipunga

+ talu < yepoko; yepoko = (succ talu)

polangipunga yepoko	31
---------------------	----

* - assumption
? - open question
+ - newly synthesized fact

Solving It - Applying New Information

tokapu alapunga talu	50
alapunga yepoko	27

+ Bigrams should be considered.

+ Juxtaposition of bigrams implies addition.

* Bigrams bind tightest to the right.

* polangipula is a typo of polangipunga

+ talu < yepoko; yepoko = (succ talu)

tokapu alapunga talu

x + y
50

polangipunga yepoko	31
---------------------	----

* - assumption
? - open question
+ - newly synthesized fact

Solving It

tokapu alapunga talu	50
alapunga yepoko	27

alapunga talu = x
alapunga yepoko = 27

- + Bigrams should be considered.
- + Juxtaposition of bigrams implies addition.
- * Bigrams bind tightest to the right.
- * polangipula is a typo of polangipunga
- + talu < yepoko; yepoko = (succ talu)

polangipunga yepoko	31
alapunga talu	26

* - assumption
? - open question
+ - newly synthesized fact

Solving It

tokapu alapunga talu	50
alapunga yepoko	27

tokapu alapunga talu

$$\begin{array}{r} x \quad + \quad 26 \\ 50 \end{array}$$

+ Bigrams should be considered.

+ Juxtaposition of bigrams implies addition.

* Bigrams bind tightest to the right.

* polangipula is a typo of polangipunga

+ talu < yepoko; yepoko = (succ talu)

polangipunga yepoko	31
alapunga talu	26

* - assumption
? - open question
+ - newly synthesized fact

Solving It- Breakthrough 1

tokapu alapunga talu	50
alapunga yepoko	27

$$\begin{array}{r} \text{tokapu alapunga talu} \\ 24 \quad + \quad 26 \\ \hline 50 \end{array}$$

+ Bigrams should be considered.

+ Juxtaposition of bigrams implies addition.

* Bigrams bind tightest to the right.

* polangipula is a typo of polangipunga

+ talu < yepoko; yepoko = (succ talu)

polangipunga yepoko	31
alapunga talu	26
tokapu	24

* - assumption
? - open question
+ - newly synthesized fact

Solving It- Breakthrough 2

tokapu talu	48
-------------	----

tokapu talu

24 × 2

48

+ Bigrams should be considered.

+ Juxtaposition of bigrams implies addition.

* Bigrams bind tightest to the right.

* polangipula is a typo of polangipunga

+ talu < yepoko; yepoko = (succ talu)

+ Juxtaposition of words in a bigram implies multiplication.

polangipunga yepoko	31
alapunga talu	26
tokapu	24
telu	1
talu	2
yepoko	3

* - assumption

? - open question

+ - newly synthesized fact

Breakthrough

Term	Number
rureponga talu	10
malapunga yepoko	15
supu	20
tokapunga telu	21
alapunga yepoko	27
polangipula talu	30
tokapu rureponga yepoko	35
tokapu malapu	40
tokapu talu	48
tokapu alapunga talu	50
tokapu talu tokapunga telu	69
tokapu talu polangipunga yepoko	79
tokapu yepoko alapunga telu	97

Number system is somewhat “positional”.

Large numbers appear to be Base-24.

Positional Numbers

四百二十

Positional Numbers

四百二十

400

20

Positional Numbers

四百二十

4

Positional Numbers

四百二十

4 100

Positional Numbers

四百二十

4 100 2

Positional Numbers

四百二十

4 100 2 10

Positional Numbers

四百二十

$$4 \times 100 + 2 \times 10$$

=

$$420$$

Solving It - The *-nga* Suffix

tokapu talu tokapunga telu

48

+

21

=

69

Solving It - The *-nga* Suffix

tokapu talu tokapunga telu

$$\begin{array}{r} 24 \quad \times \quad 2 \quad + \quad 21 \\ = \\ 69 \end{array}$$

Solving It - The *-nga* Suffix

tokapu talu tokapunga telu

$$\begin{array}{cccccccc} 24 & \times & 2 & + & 24 & ? & \times & 1 \\ & & & & & = & & \\ & & & & & 69 & & \end{array}$$

Solving It- The *-nga* Suffix

rureponga talu	10
malapunga yepoko	15
tokapunga telu	21
alapunga yepoko	27

$$\begin{aligned} \underline{\text{rureponga talu}} &= 10 \\ 12 \quad ? \quad 2 &= 10 \\ \underline{\text{malapunga yepoko}} &= 15 \\ 16 \quad ? \quad 3 &= 15 \\ \underline{\text{tokapunga telu}} &= 21 \\ 24 \quad ? \quad 1 &= 21 \\ \underline{\text{alapunga yepoko}} &= 27 \\ x \quad ? \quad 3 &= 27 \end{aligned}$$

+ Bigrams should be considered.

+ Juxtaposition of bigrams implies addition.

* Bigrams bind tightest to the right.

* polangipula is a typo of polangipunga

+ talu < yepoko; yepoko = (succ talu)

+ Juxtaposition of words in a bigram implies multiplication.

* - assumption
? - open question
+ - newly synthesized fact

Solving It- The *-nga* Suffix

rureponga talu	10
malapunga yepoko	15
tokapunga telu	21
alapunga yepoko	27

$$\underline{\text{rureponga talu}} = 10$$

$$12 \quad -4 \quad 2 = 10$$

$$\underline{\text{malapunga yepoko}} = 15$$

$$16 \quad -4 \quad 3 = 15$$

$$\underline{\text{tokapunga telu}} = 21$$

$$24 \quad -4 \quad 1 = 21$$

$$\underline{\text{alapunga yepoko}} = 27$$

$$28 \quad -4 \quad 3 = 27$$

+ Bigrams should be considered.

+ Juxtaposition of bigrams implies addition.

* Bigrams bind tightest to the right.

* polangipula is a typo of polangipunga

+ talu < yepoko; yepoko = (succ talu)

+ Juxtaposition of words in a bigram implies multiplication, except following *-nga*, then it's addition.

+ *-nga* means (-4).

* - assumption

? - open question

+ - newly synthesized fact

Solving It - The *-nga* Suffix

tokapu talu tokapunga telu

$$\begin{array}{ccccccccc} 24 & \times & 2 & + & 24 & ? & \times & 1 & \\ & & & & & & & & \\ & & & = & & & & & \\ & & & & 69 & & & & \end{array}$$

Solving It - The *-nga* Suffix

tokapu talu tokapunga telu

$$\begin{array}{r} 24 \times 2 + 24 + -4 + 1 \\ = \\ 69 \end{array}$$

Inconsistencies in Positional Numbers

69420

sixty nine thousand four hundred and twenty

$$(6 \times 10 + 9) \times 1000 + 4 \times 100 + 2 \times 10$$

六万九千四百二十

$$6 \times 10000 + 9 \times 1000 + 4 \times 100 + 2 \times 10$$

Solving It...

Left as an exercise to the watcher.

ML/AI = “Do What You Just Did, At Scale”

What The Human Needed

- Which examples to work on
- Recursive problem solving (solving for something while solving for another)
- Parallel problem solving (solving for multiple things at once)
- Backtracking
- Error correction
- Feature engineering
- Prior knowledge
 - Arithmetics
 - Algebra
 - Linguistics
 - Statistics
- Putting all these together

What Machines Are Good At

- Which examples to work on
- Recursive problem solving (solving for something while solving for another)
- Parallel problem solving (solving for multiple things at once)
- Backtracking
- Feature engineering
- Error tolerance
- Prior knowledge*
- Putting all these together

AI Has a Long Way to Go

Human reasoning is still needed.

Reinforcement learning *may* learn “reasoning”.

Telu More Experiment

Inspecting the BERT Neurons

“fourty eight” → “4 8”

“tokapu talu” → “4 8”



Layer 11, Head 5 (EN), Head 6 (UU)

Talu Artificial Languages

Base-10: One byte for units under 5, one byte-pair for units up to 10, multiply-add combinations for the rest up to 100

Base-12: One byte for units under 6, one byte-pair for units up to 12, multiply-add combinations for the rest up to 100.

Two ways of doing multiply-add: prefix and postfix multiplication

The Basic Components

Base-10

a	1	ba	6
e	2	be	7
i	3	bi	8
o	4	bo	9
u	5	bu	10

Base-12

a	1	ba	7
e	2	be	8
i	3	bi	9
o	4	bo	10
u	5	bu	11
ə	6	bə	12

Examples 1 - Postfix Multiplication

Base-10

abu = 11

ebu = 12

bue = 20

abue = 21

obue = 24

ebube = 72

Base-12

abe = 13

ebe = 14

bee = 24

abea = 25

ebea = 26

beə = 72

Examples 2 - Prefix Multiplication

Base-10

bua = 11

bue = 12

ebu = 20

ebua = 21

ebuo = 24

bebue = 72

Base-12

bea = 13

bee = 14

ebe = 24

ebea = 25

ebee = 26

əbe = 72

Can a BERT-based LM Translate These Artificial Languages?

Multiply-Add Type	Base-10	Base-12
Prefix multiplication (e.g. “ <i>twenty-four</i> ”)	Yes	No
Postfix multiplication (e.g. “ <i>four-and-twenty</i> ”)	No	No

Can a LM w/ BERT Arch Translate These Artificial Languages?

Multiply-Add Type	Base-10	Base-12
Prefix multiplication (e.g. “ <i>twenty-four</i> ”)	Yes*	Yes*
Postfix multiplication (e.g. “ <i>four-and-twenty</i> ”)	Yes*	Yes*

* super over-fitted obviously

Yepoko Lessons

Use The Right Tool for the Right Job

Machine learning algorithms are probably not the right tool for this puzzle.

It's the right tool for a much larger dataset.

Prolog might help.

Careful Thought with Judiciously Placed Statistical Tools

For now, we can't replace careful thought with machines.

Machines are awesome at statistics though.

Watch Your Biases!

Bias can fuck you up in more ways than you expect.

Closing Thoughts

The Language

- Umbu-ungu/Imbo-ungu is a language in Southern Highlands of PNG.
- Base-4, Base-12, Base-24, Base-28, Base-32 number system.
- PNG is the most linguistically diverse country in the world.

Neural Machine Translation

- Requires parallel corpus.
- Only parallel corpus for Umbu-Ungu is The Bible.
- Rare language communities are underserved.

Join Me (Shameless Plug)

<https://gorgonia.org>

chewxy@gmail.com

 @chewxy

Thank You

The Fate of Umbu-Ungu

andrete - 100 (from *hundred* in English)

tausen - 1000 (from *thousand* in English)

Mostly Tok Pisin (from talk business)

The Deep Learning Bombe - An Enigma