เพิ่มชื่อบริษัท

Bi-Weekly Progress Report

Word Correction

29 March 2023

Exploring Data

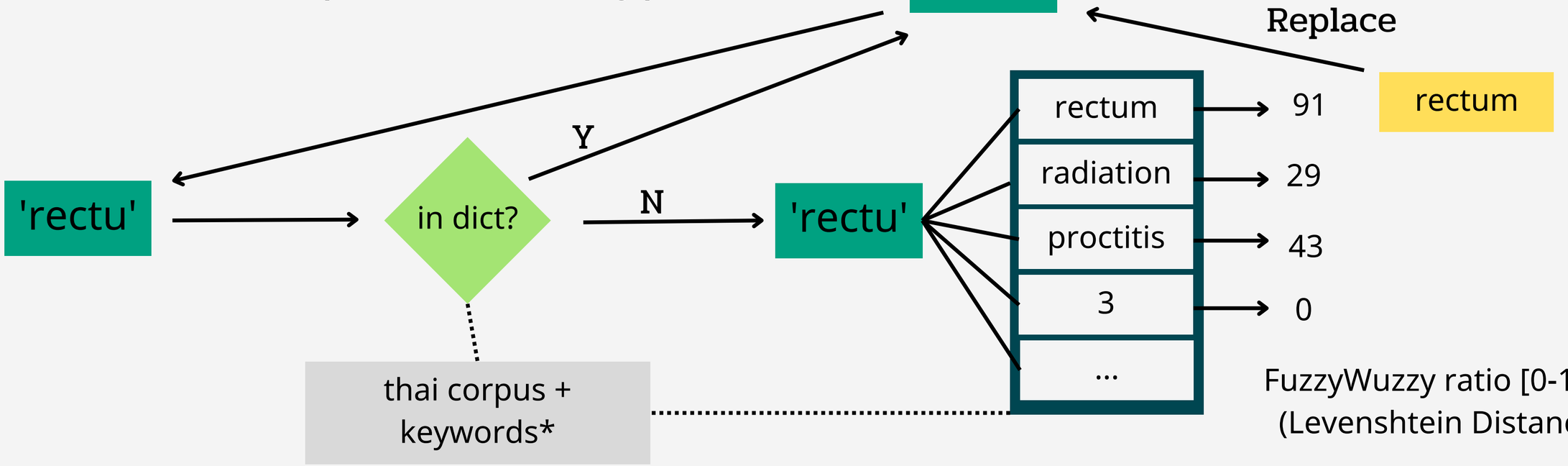
ID	Error case	Example
0	word error	• Transverse: ถ, ตานเมตร, ทันเมตร, ตาลนิเมตร, ตาลนวิส, ตาลิฟติ, ตาน worth, 1 ถานเมตร, ταν̆วเมตร • Sysndome : ˘ตร , ภินติ, sz̄นติum, สินติum, sz̄นติum
1	Edit distance	Proctitis -> Pocticis
2	Real word error (คำถูกต้อง แต่ไม่ถูกบริบท)	• ใหญ่: ยาย, ย่าย • ใส่ตัวเลขมาผิด จาก target ที่ต้องการ • serrated: Solitary
3	Missing data	index(5944)

1st Approach: Naive approach

"Radiation proctitis Type 3 Rectu"

- Lower case
- word_tokenize by pythainlp (newmm)

['radiation', ' ', 'proctitis', ' ', 'type', ' ', '3', ' ', 'rectu']



1st: Result

```
start = time.time()
trans1 = "Hello DeepGI Submucodddsal Rweation Pocticis เล็กกว่า 5 มิลลิเมตร NBI Descding Toctitis"
target = "Hello DeepGI Submucosal Radiation proctitis เล็กกว่า 5 มิลลิเมตร NBI Descending proctitis"
corrected_trans = correct_sentence(trans1, vocabs)

end = time.time()

print("Process time:", end - start)
corrected_trans == target.lower(), corrected_trans
```

```
Process time: 5.282670497894287
(True,
 'hello deepgi submucosal radiation proctitis เล็กกว่า 5 มิลลิเมตร nbi descending proctitis')
```

- ใช้เวลาในการประมวลผลนาน
- ทำได้เฉพาะ word ที่เป็นการ edit operation (insert, delete, replace) ส่งผลให้แก้พวกคำไทยไม่ได้
- ไม่ได้ดู context การการแก้ไข

1st: How to improve

- ใช้เวลาในการประมวลผลนาน
  - สามารถแก้ได้โดยใช้ SymSpell โดยการเพิ่มขนาดข้อมูล generate คำที่เป็นไปได้ แล้วพิจารณาเฉพาะ delete operation
- ทำได้เฉพาะ word ที่เป็นการ edit operation (insert, delete, replace) ส่งผลให้แก้พวกคำไทยไม่ได้
  - ยังแก้ไม่ได้
- ไม่ได้ดู context การการแก้ไข
  - สามารถที่จะใช้ Norvig's apporch ได้ ที่มีการเพิ่ม language model (unigram) เข้ามา แต่ใน implementation [https://norvig.com/spell-correct.html] ตรงส่วน Error model ใช้การจัดลำดับเทียบกับ candidate ที่สร้างขึ้นมาจาก edit distance ทำให้สุดท้ายก็ยังแก้ปัญหา context ยังไม่ได้ และรองรับแค่ edit distance
  - อีกทั้ง language model ถ้าจะมาใช้กับโจทย์ปัจจุบัน ไม่รู้จะทำไง

เสริมใน link มีการ optimize symspell ให้ลดขนาดข้อมูลลง
https://towardsdatascience.com/spelling-correction-how-to-make-an-accurate-and-fast-corrector-dc6d0bcbba5f

Exploring Data

ID	Error case	Example
0	word error	• Transverse: ถ, ตานเมตร, ทันเมตร, ตาลนิเมตร, ตาลนวิส, ตาลิฟติ, ตาน worth, 1 ถานเมตร, ταν̆วเมตร • Sysndome : ˘ตร❓ , ภินติ, sz̄นติum, สินติum, sz̄นติum
1	Edit distance	Proctitis -> Pocticis
2	Real word error (คำถูกต้อง แต่ไม่ถูกบริบท)	• ใหญ่: ยาย, ย่าย • ใส่ตัวเลขมาผิด จาก target ที่ต้องการ • serrated: Solitary
3	Missing data	index(5944)

ID 2: Real word error

Paper: https://aclanthology.org/O13-1022.pdf
word is meaningful but not the intended word in the context of the sentence.

"Radiation proctitis Type 3 Rectu"

['type', ' ty', 'tp', 'tpe ', ...]

$\mathbf{w}_0^i \quad \mathbf{w}_1^i \quad \mathbf{w}_2^i \quad \mathbf{w}_3^i$

$$Score\left(W_j^i\right) = P_1\left(W_j^i|W^{i-1}\right) + P_2\left(W_j^i|W^{i+1}\right) + P_3\left(W_j^i|W^{i-1},W^{i+1}\right)$$

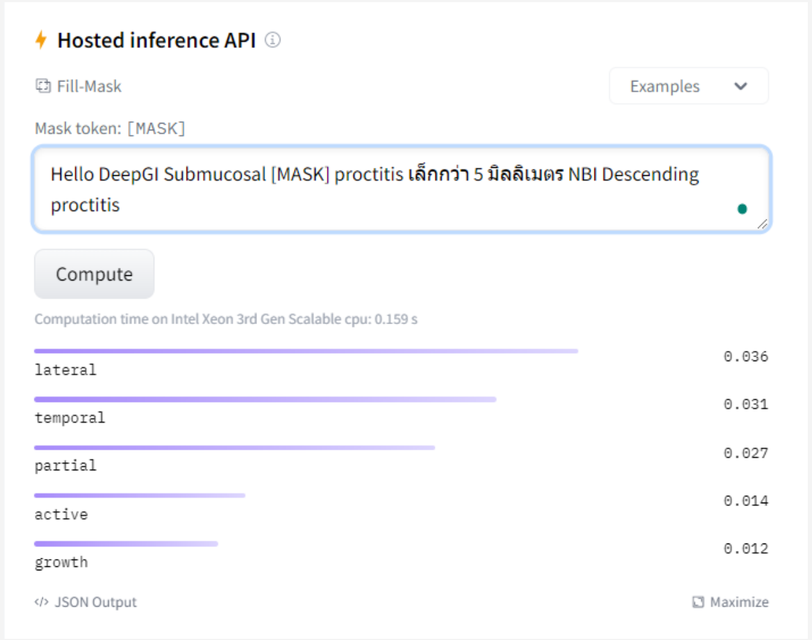
$$Score\left(W_j^i\right) = \lambda_1 P_1\left(W_j^i|W^{i-1}\right) + \lambda_2 P_2\left(W_j^i|W^{i+1}\right) + \lambda_3 P_3\left(W_j^i|W^{i-1},W^{i+1}\right)$$

!!Require Language model

ID 3: Non-word error**

source:
https://www.researchgate.net/publication/353222184_Generating_Correction_Candidates_for_OCR_Errors_using_BERT_Language_Model_and_FastText_SubWord_Embeddings

Post-processing: Detecting the errors, generating the candidates, correcting the errors



BERT

"verj^"

↓

"ver", "verj", "jerk", "very", ...

FastText

เพิ่มชื่อบริษัท

Bi-Weekly Progress Report

Word Correction

12 Apil 2023

Exploring data (model 2*)

ID	Error case	Rate	Description
1	Word error	0.4799	คำที่ไม่มีใน dict
2	Edit distance	part of word error (0.4799)	Proctitis -> Pocticis
3	Wrong word	0.064	เป็นคำที่มีใน dict แต่ บริบท หรือตำแหน่งไม่ถูกต้อง
4	Missing data	0.1495	token ที่ได้หลังจาก tokenize ของ reference กับ transcript ไม่เท่ากัน

Accuracy (exact match): 0.452
Wer score: 0.118

Note:

- Dictionary = Keywords + set(tokenized(reference))
- edit distance $\subseteq$ word error

Some errors (model 2*)

- Word error
  - proctitis: toctasis, toctitis
  - type: tidene, tile, tine
  - flat: fat
- Wrong word
  - **ตัวเลขผิด**
  - hello deepgi 10.5 **มิลลิเมตร** sessile polyp nbi benign ascending esd > hello deepgi 5.5 **มิลลิเมตร** sessile polyp nbi benign ascending esd

Extreme cases (model 2*)

target_text	input_text
hello deepgi 1 เซนติเมตร flat subepithelial type 3 hyperplastic polyp anus emr	hello deepgi 1 เซนติเมตร flat submitting ฟีตร submitting ฟีตร submitting ฟีตร submitting ฟีตร submitting ฟีตร submitting ฟีตร submitting ฟีตร submitting ฟีตร submitting ฟีตร submitting ฟีตร submitting ฟีตร submitting ฟีตร submittin submitting ฟีตร submitting ฟีตร submitting ฟีตร submitting ฟีตร submitting ฟีตร submitting ฟีตร submitting ฟีตร submitting ฟีตร submitting ฟีตร submitting ฟีตร submitting ฟีตร submitting ฟีตร submitting ฟีตร submi ฟีตร submitting ฟีตร submitting ฟีตร submitting ฟีตร submitting ฟีตร submitting ฟีตร submitting ฟีตร submitting ฟีตร submitting ฟีตร submitting ฟีตร submitting ฟีตร submitting ฟีตร submitting ฟีตร
hello deepgi เล็กกว่า 5 มิลลิเมตร sessile cancer type 3 flat polyp anus biopsy	hello deepgi ɲtɥกว่า 5 มิลลิเมตร sessile cancer ɡnn
hello deepgi 1 เซนติเมตร pedunculated solitary rectal ulcer syndrome type 3 lipoma anus emr	hello deepgi 1 เซนติเมตร pedaculated solitary rectal ulcer ดินดิ ด
hello deepgi 10.5 มิลลิเมตร flat hemorrhoid type 3 radiation proctitis anus biopsy	hello deepgi 5.5 มิลลิเมตร fat hemorrhoid ถ่าย 3 ริディ เมตร proctitis anus biospy

Norvig approach

- **รองรับ** error **ที่เกิดขึ้นได้มากสุด** 2 operation(edit, remove, add, transposes)
- Dictionary = Keywords + set(tokenized(reference) (84 words)

```
Before: hello deepgi เล็กกว่า 5 มิลลิเมตร sessile polyp  type 3 flat polyp anus biospy
After: hello deepgi เล็กกว่า 5 มิลลิเมตร sessile polyp type 3 flat polyp anus biopsy
Time: 0.0014331340789794922
```

```
Before: hello deepgi 5.5 มิลลิเมตร sessile cancer nbi radiation ptoctitis ascending sampling
After: hello deepgi 5.5 มิลลิเมตร sessile cancer nbi radiation proctitis ascending sampling
Time: 0.0019512176513671875
```

```
Before: hello deepgi 10.5 มิลลิเมตร sessile cancer thai 3 pseudopolyp anus emr
After: hello deepgi 10.5 มิลลิเมตร sessile cancer thai 3 pseudopolyp anus emr
Time: 0.2793216705322265 6
```

Exploring data after correction (norvig)

ID	Error case	Rate	New rate Norvig	Description
1	Word error	0.4799	0.3 (-0.18)	คำที่ไม่มีใน dict
2	Edit distance	part of word error	-	Proctitis -> Pocticis
3	Wrong word	0.064	0.104(+0.04)*	เป็นคำที่มีใน dict แต่ บริบท หรือตำแหน่งไม่ถูกต้อง
4	Missing data	0.1495	0.1495	token ที่ได้หลังจาก tokenize ของ reference กับ transcript ไม่เท่ากัน

Accuracy (exact match): 0.452 -> 0.586 (*ตัวที่ถูกอยู่แล้วไม่ลดลง)
Wer score: 0.118 -> 0.098

Dictionary = Keywords + set(tokenized(reference))

Seq2Seq model

- Additive attention model (class NLP) with 50 epoch

```
EXAMPLES = ['hello deepgi ascending abnormal vascular 10.5 มิลลิเมตร sessile serrated adenoma type 3',
            "hello deepgi 10.5 มิลลิเมตร sessile polyp ถ่าย 3 เสือ pseudopolyp anus biospy"]
INFO:pytorch_lightning.accelerators.cuda:LOCAL_RANK: 0 - CUDA_VISIBLE_DEVICES: [0]
Predicting DataLoader 0: 100% 2/2 [
hello deepgi 10.5 มิลลิเมตร มิลลิเมตร abnormal vascular type 3 radiation proctitis ascending biopsy<EOS><PAD><PAD><PAD><PAD><PAD><PAD><PAD><PAD><PAD><PAD><PAD><PAD><PAD><PAD>
hello deepgi 10.5 มิลลิเมตร sessile polyp type 3 pseudopolyp anus biopsy<EOS><PAD><PAD><PAD><PAD><PAD><PAD><PAD><PAD><PAD><PAD><PAD><PAD><PAD><PAD><PAD><PAD><PAD>
[None, None]
```

- **BART model (facebook/m2m100_418M)

```
input: hello deepgi 10.5 มิลลิเมตร sessile cancer ถ่าย 3 มิลลิเมตร pseudopolyp anus biosp
target: ro_RO hello deepgi 10.5 มิลลิเมตร sessile cancer type 3 pseudopolyp anus biopsy

input: hello deepgi 10.5 มิลลิเมตร sessile cancer thai 3 มิลลิเมตร pseudopolyp anus biossp
target: ro_RO hello deepgi 10.5 มิลลิเมตร sessile cancer type 3 pseudopolyp anus biopsy

input: hello deepgi rectu solitary rectal ulcer synซ้ome เล
target: ro_RO hello deepgi เล็กกว่า 5 มิลลิเมตร flat hemunculated solitary rectal ulcer syndrome type 1 flat polyp sigmoid cold snare

input: hello deepgi rectu soldffitary rectal ulcer syndrome
target: ro_RO hello deepgi 10.5 มิลลิเมตร flat circumferential rectal ulcer syndrome type 1 benign sigmoid polypectomy

input: hello deepgi 10.5 มิลลิเมตร sessile polyp ถ่าย 3 เสือ pseudopolyp anus biospy
target: ro_RO hello deepgi 10.5 มิลลิเมตร sessile polyp type 3 pseudopolyp anus biopsy

input: hello deepgi 1 เซนติเมตร flat submitting ฟัตร submitting ฮิลลิว nbi hyperplastic polyp ascending esd
target: ro_RO hello deepgi 1 เซนติเมตร flat subepithelial nbi hyperplastic polyp ascending esd
```

Exploring data after correction (encode-decode)

ID	Error case	Rate	New rate Norvig	New Rate encode-decode	Description
1	Word error	0.4799	0.3	0.0	คำที่ไม่มีใน dict
2	Edit distance	part of word error	-	-	Proctitis -> Pocticis
3	Wrong word	0.064	0.104	0.035	เป็นคำที่มีใน dict แต่ บริบท หรือตำแหน่งไม่ถูกต้อง
4	Missing data	0.1495	0.1495	0.025	token ที่ได้หลังจาก tokenize ของ reference กับ transcript ไม่เท่ากัน

Accuracy (exact match): 0.452 -> 0.586 -> 0.962
Wer score: 0.118 -> 0.098 -> 0.0095

What about the rest of errors?

Accuracy (exact match): 0.452 -> 0.586 -> 0.962
Wer score: 0.118 -> 0.098 -> 0.0095

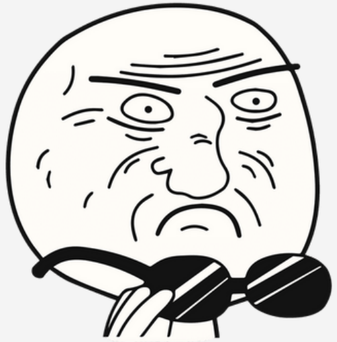
550	hello deepgi 1 เซนติเมตร flat subepithelial type 3 solitary rectal ulcer syndrome anus emr	hello deepgi 1 เซนติเมตร flat submitting ฟีตร

1794	hello deepgi 1 เซนติเมตร pedunculated solitary rectal ulcer syndrome type 3 lipoma anus emr	hello deepgi 1 เซนติเมตร pedaculated solitary rectal ulcer ตินติ ต
432	hello deepgi 1 เซนติเมตร pedunculated solitary rectal ulcer syndrome nbi lipoma ascending sampling	hello deepgi 1 เซนติเมตร pedacurate solitary rectal ulcer ตินติ

153	hello deepgi 10.5 มิลลิเมตร pedunculated solitary rectal ulcer syndrome type 1 benign sigmoid cold snare	hello deepgi 5 มิลลิเมตร pedacurate solitary rectal ulcer ตินติ	hello deepgi 10.5 มิลลิเมตร pedunculated solitary rectal ulcer syndrome type 1 benign sigmoid polypectomy
261	hello deepgi 10.5 มิลลิเมตร flat subepithelial type 3 benign anus biopsy	hello deepgi 5 มิลลิเมตร flat subbit มิลลิเมตร	hello deepgi 10.5 มิลลิเมตร flat subepithelial type 3 benign anus emr
284	hello deepgi 10.5 มิลลิเมตร flat circumferential mass type 3 benign anus emr	hello deepgi 10.5 มิลลิเมตร flat circumferential mass เซิรควิตร	hello deepgi 10.5 มิลลิเมตร flat circumferential mass type 3 benign anus biopsy
287	hello deepgi 10.5 มิลลิเมตร flat circumferential mass type 3 radiation proctitis anus biopsy	hello deepgi 10.5 มิลลิเมตร flat circumferential mass เซิรควิตร เซิรควิตร เซิร	hello deepgi 10.5 มิลลิเมตร flat circumferential mass type 3 benign anus biopsy
342	hello deepgi 1 เซนติเมตร sessile lipoma type 1 lipoma sigmoid polypectomy	hello deepgi 1 เซนติเมตร sessile lipoma type 1 lipoma sigmoid polypectomy	hello deepgi 1 เซนติเมตร sessile polyp type 1 lipoma sigmoid polypectomy

!!! Wait a minute

- Wrong word
  - ตัวเลขผิด
  - hello deepgi 10.5 มิลลิเมตร sessile polyp nbi benign ascending esd > hello deepgi 5.5 มิลลิเมตร sessile polyp nbi benign ascending esd



Should we correct it or not?

2	hello deepgi 10.5 มิลลิเมตร sessile polyp nbi benign ascending sampling	hello deepgi 5.5 มิลลิเมตร sessile polyp nbi benign ascending sampling	hello deepgi 10.5 มิลลิเมตร sessile polyp nbi benign ascending sampling
3	hello deepgi 10.5 มิลลิเมตร sessile polyp nbi benign ascending esd	hello deepgi 5.5 มิลลิเมตร sessile polyp nbi benign ascending esd	hello deepgi 10.5 มิลลิเมตร sessile polyp nbi benign ascending esd
4	hello deepgi 10.5 มิลลิเมตร sessile polyp nbi radiation proctitis ascending sampling	hello deepgi 5.5 มิลลิเมตร sessile polyp nbi radiation proctitis ascending sampling	hello deepgi 10.5 มิลลิเมตร sessile polyp nbi radiation proctitis ascending sampling

!!Overfit

```
input: rectu solitary
 result: ro_RO hello deepgi 10.5 มิลลิเมตร flat subepithelial type 1 solitary rectal ulcer syndrome sigmoid cold snare

input: pour about preparation with solid stone throughout the right column
 result: ro_RO hello deepgi เล็กกว่า 5 มิลลิเมตร flat colitis type 3 flat polyp anus biopsy

input: hot snare polypectomy
 result: ro_RO hello deepgi 10.5 มิลลิเมตร flat subepithelial type 1 malignant non invasive polyp sigmoid polypectomy

input: flat polp 3 มิลลิเมตร ทิตรันสเมตร
 result: ro_RO hello deepgi 10.5 มิลลิเมตร flat subepithelial type 3 pseudopolyp anus biopsy
```

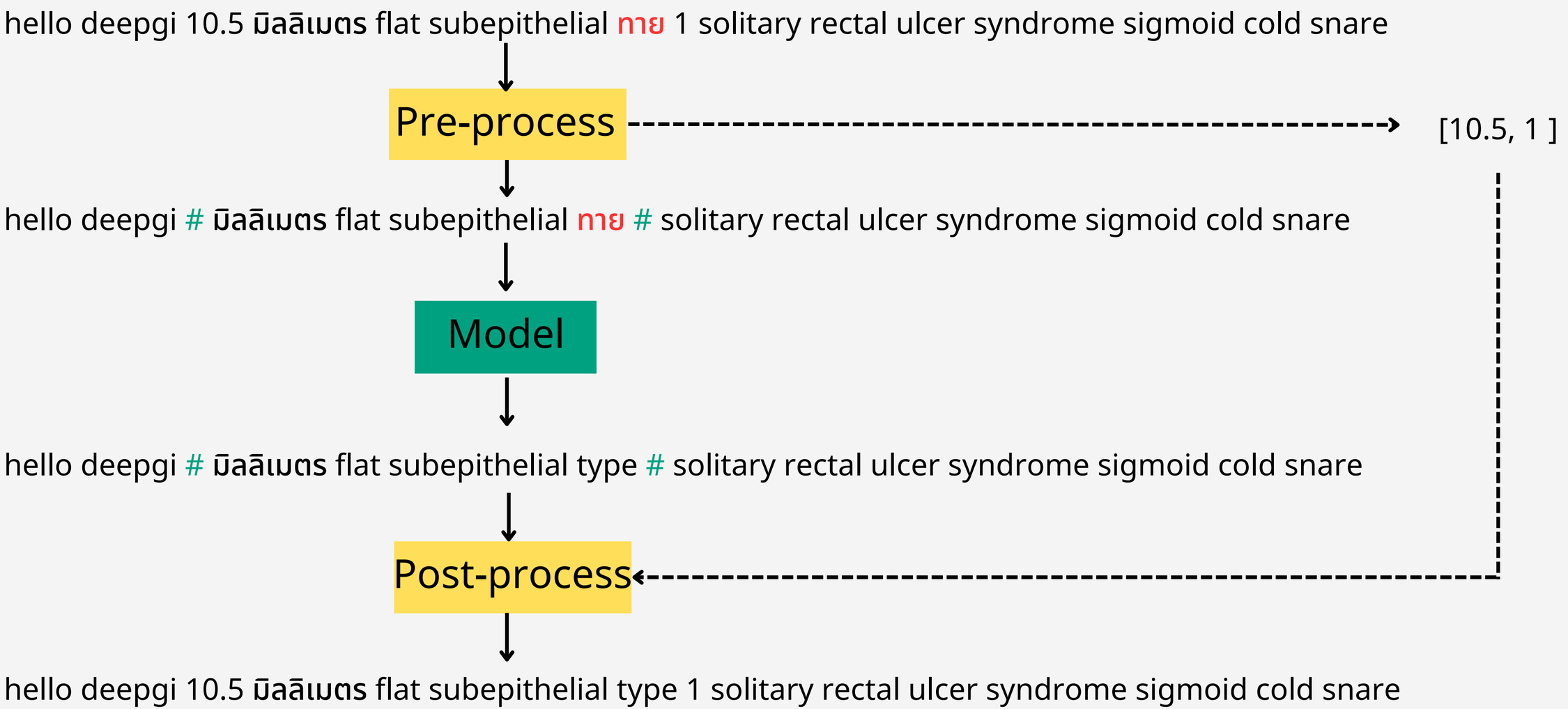
เพิ่มชื่อบริษัท

Bi-Weekly Progress Report

Word Correction

4 March 2023

Preprocess and Postprocess Text



เพิ่มชื่อบริษัท

Final Report

Word Correction

18 March 2023

Approaches

Norvig approach

- รองรับ error ที่เกิดขึ้นได้มากสุด 2 operation(edit, remove, add, transposes)
- Dictionary = Keywords + set(tokenized(reference) (84 words)

Bart model

Pair: transcribe <-> reference

Bart+Mask_number model

Pair: transcribe <-> reference

hello deepgi 10.5 มิลลิเมตร flat subepithelial ทาย 1 solitary rectal ulcer syndrome sigmoid cold snare

↓

Pre-process

↓

hello deepgi # มิลลิเมตร flat subepithelial ทาย # solitary rectal ulcer syndrome sigmoid cold snare

Results of Models

Evaluate metric	Raw Data	Norvig	Bart	Bart+mask_Num	Description
word_error_rate	**48%**	**30%**	**0%**	**8%**	The word is not in "Custom dict"
wrong_number_rate	**20%**	20%	**2%**	20%	The transcript numbers don't match the reference numbers.
exceed_data_rate	**8%**	8%	**2%**	**3%**	The length of the trans text > the ref's length.
missing_data_rate	**7%**	7%	**1%**	**0%**	The length of the trans text < the ref's length.
wrong_word_rate	**6%**	**10%**	**4%**	**14%**	The corrected words in the trans text are not existed in the ref text.
wer_score	**12%**	**10%**	**1%**	**3%**	WER score
accuracy	**45%**	**59%**	***96%**	**78%**	Exact match (the reference text === the transcript text)

Refactor code

link: https://colab.research.google.com/drive/1U6d0ud93jK5INk_nix8QqAbZsEynaL0C?usp=sharing

Class DeepGICorrector()

ซ่อน 1 เซลล์

Demo

```
[ ]  sentence = df['input_text'][42]
     corrector = DeepGICorrector(custom_dict, lower_case=True)

     start = time.time()

     print("Before:", sentence)
     print("After:",corrector.correct(sentence))
     end = time.time()

     print("Time:",end - start)
```

```
Before: Hello DeepGI 10.5 มิลลิเมตร Sessile Cancer Thai 3 Pseudopolyp Anus EMR
After: hello deepgi 10.5 มิลลิเมตร sessile cancer thai 3 pseudopolyp anus emr
Time: 0.10617876052856445
```

Refactor code

link: https://colab.research.google.com/drive/1U6d0ud93jK5INk_nix8QqAbZsEynaL0C?usp=sharing

Class DeepGIBartCorrection()

ซ่อน 1 เซลล์

Demo

```
[55] # initial model
     corrector = DeepGIBartCorrection(epochs=3, max_length=60, batch_size=8, mask_num=True)

     # Train model
     corrector.fit(train, test)


     # model prediction
     samples = [
             "hello deepgi 10.5 มิลลิเมตร sessile cancer ถ่าย 3 มิลลิเมตร pseudopolyp anus biosp".lower(),
             "hello deepgi 10.5 มิลลิเมตร sessile cancer thai 3 มิลลิเมตร pseudopolyp anus biosssp".lower(),
              "Hello DeepGI Rectu Solitary rectal ulcer synซ้ome เล".lower(),
              "Hello DeepGI Rectu Soldffitary rectal ulcer syndrome ".lower(),
              'hello deepgi 5.5 มิลลิเมตร sessile polyp ถ่าย 3 เสื่อ pseudopolyp anus biospy'.lower(),
              'hello deepgi 1 เซนติเมตร flat submitting ฟัตร submitting ฮิลลิว nbi hyperplastic polyp ascending esd'.lower(),
              'Rectu solitary'.lower(),
              'pour about preparation with solid stone throughout the right column'.lower(),
              'hot snare polypectomy'.lower(),
              'flat polp 3 มิลลิเมตร ทิตรันสเมตร'.lower()
             ]

     predicts = corrector.predict(samples)
     for i,s in enumerate(samples) :
       print(f"input: {s} \nresult: {predicts[i]}\n")
```