Contextual Vietnamese Spelling Correction

Đỗ Anh Khoa - 1852471

Nguyễn Trần Hiếu - 1852370

Table Of Content

01

What & Why Correct Spelling?

Introduction

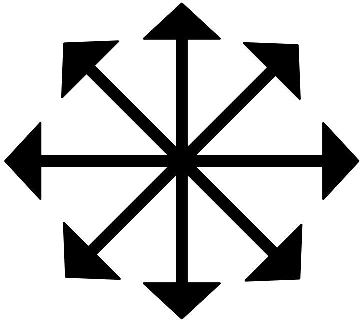
Given a text, **identify** misspelling or misplaced words and **suggest** alternatives within the **context** of a sentence.

EXAMPLE:

1. Trên đường đi hocj, ban An nhặt được tiền rơi.
2. Những nguyên tắc cơ bản trong việc ba n hành sửa đổi hiến phápdduowcjc thực thi
3. Cây bàn cao lớn như một vệ sĩ lặng lex âm thầm.

Why?

Document, mail,
content crafting.



Expanded to other
field

Search engines,
Building application

Error Types

ERROR TYPES	DEFINITIONS	EXAMPLES
Typos	Errors generated by typing text such as character insertion, deletion, replace close-proximity character; typing mechanism like telex, VNI; inappropriate white-space insert or deletion.	- “Uống’ → “Uuongs” - “Ăn” → “A8n” - “Đi ăn” → “điăn” - “Đi” → “Dsi”
Spelling Errors	Unofficial conventions or regional dialects in Vietnam, or misused knowledge.	- “Không’ → “Hông” - “Dang rộng” → “Giang rộng” - “Cá cược” → “Cá cượt”
Diacritic Errors	Vowels with missing or wrong dialect.	- “Mỗi ngày’ → “Moi ngay” - “Đầy ắp“ → “Đầy ấp”

Error Types

Out-of-scope cases

Error Types

ERROR TYPES

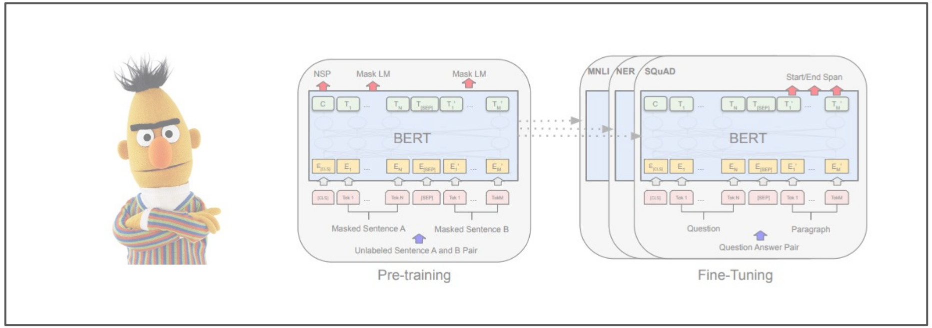
- **Typos**: errors generated by typing text such as character insertion, deletion, replace close-proximity character; typing mechanism like telex, VNI; inappropriate white-space insert or deletion.
- **Spelling Errors:** due to unofficial conventions or regional dialects in Vietnam, or misused knowledge.
- **Diacritic Errors**: vowels with missing or mistaken, wrong dialect.

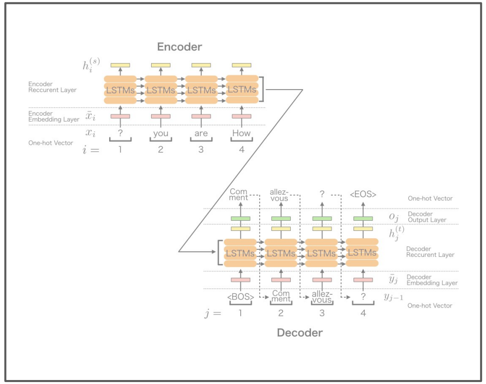
Examples

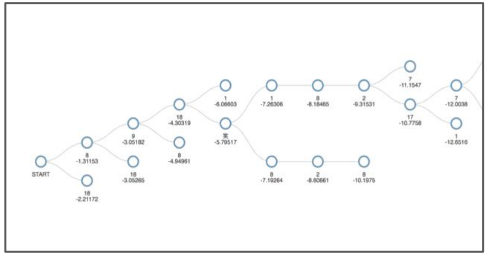
- **Typos:**
    - "Uống' → "Uuongs"
    - "Ăn" → "A8n"
    - "Đi ăn" → "điăn"
- **Spelling Errors:**
    -

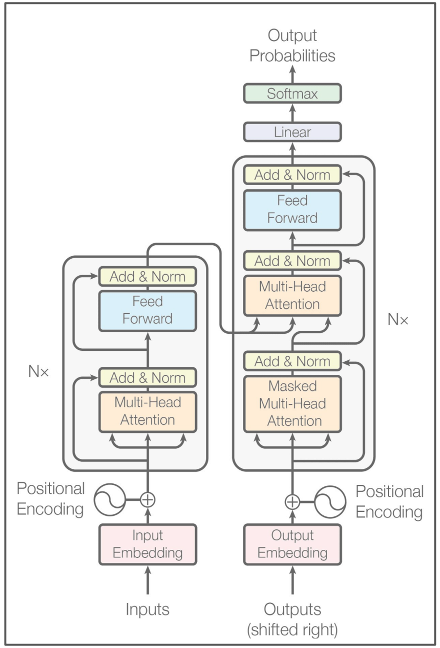
02

Background Works









Background Works

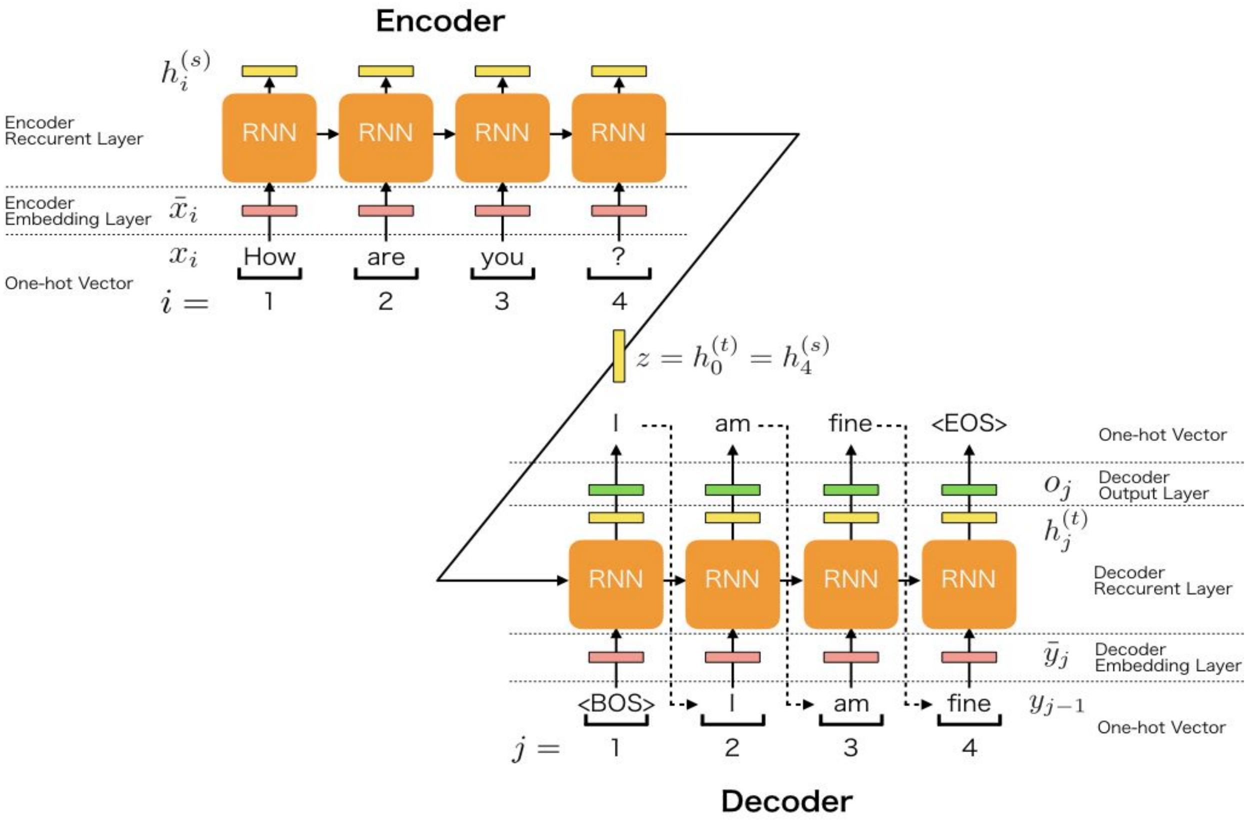


Fig 1. Sequence to sequence model with RNN

Background Works

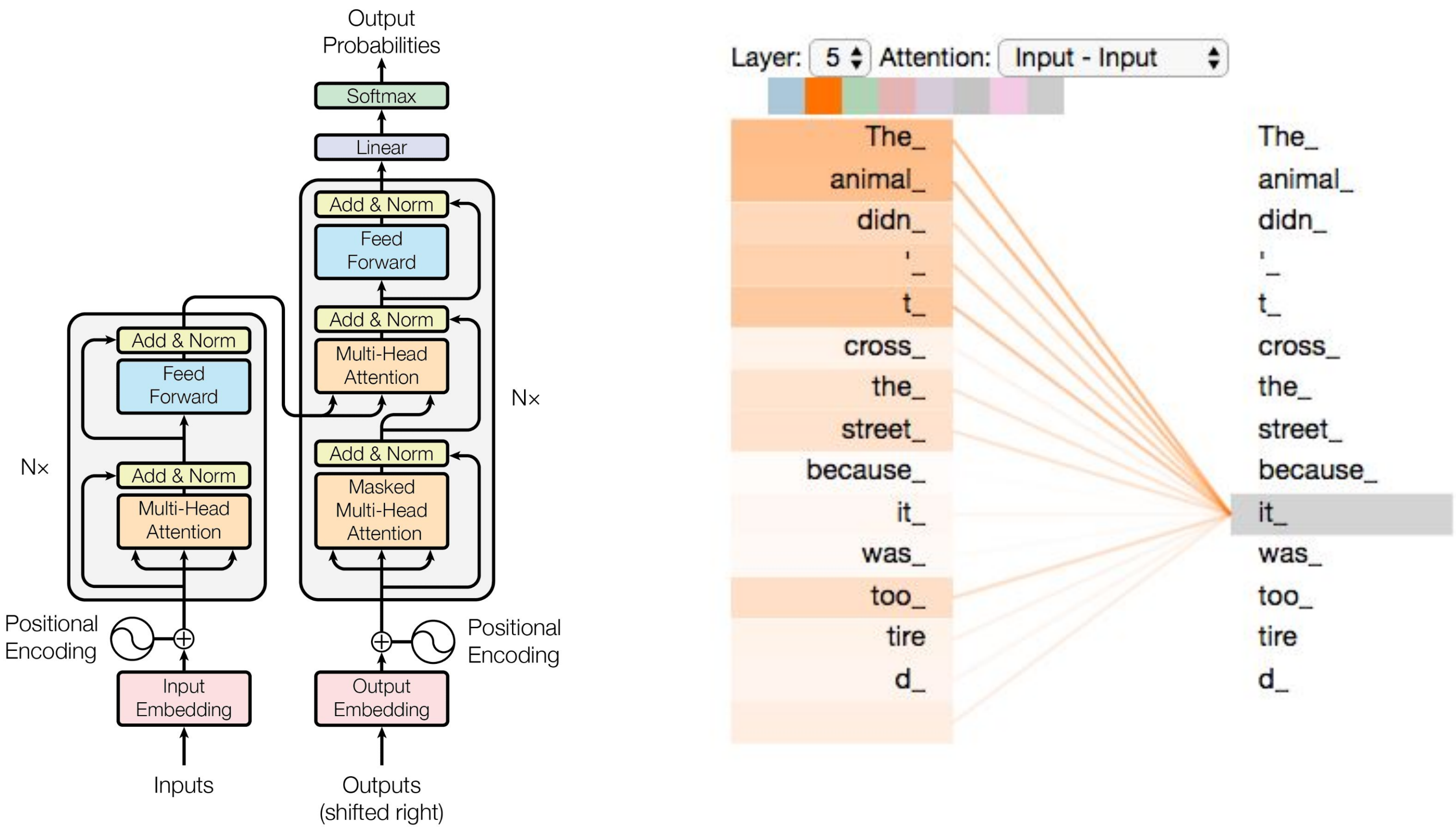


Fig 2. Transformer Architecture

Background Works

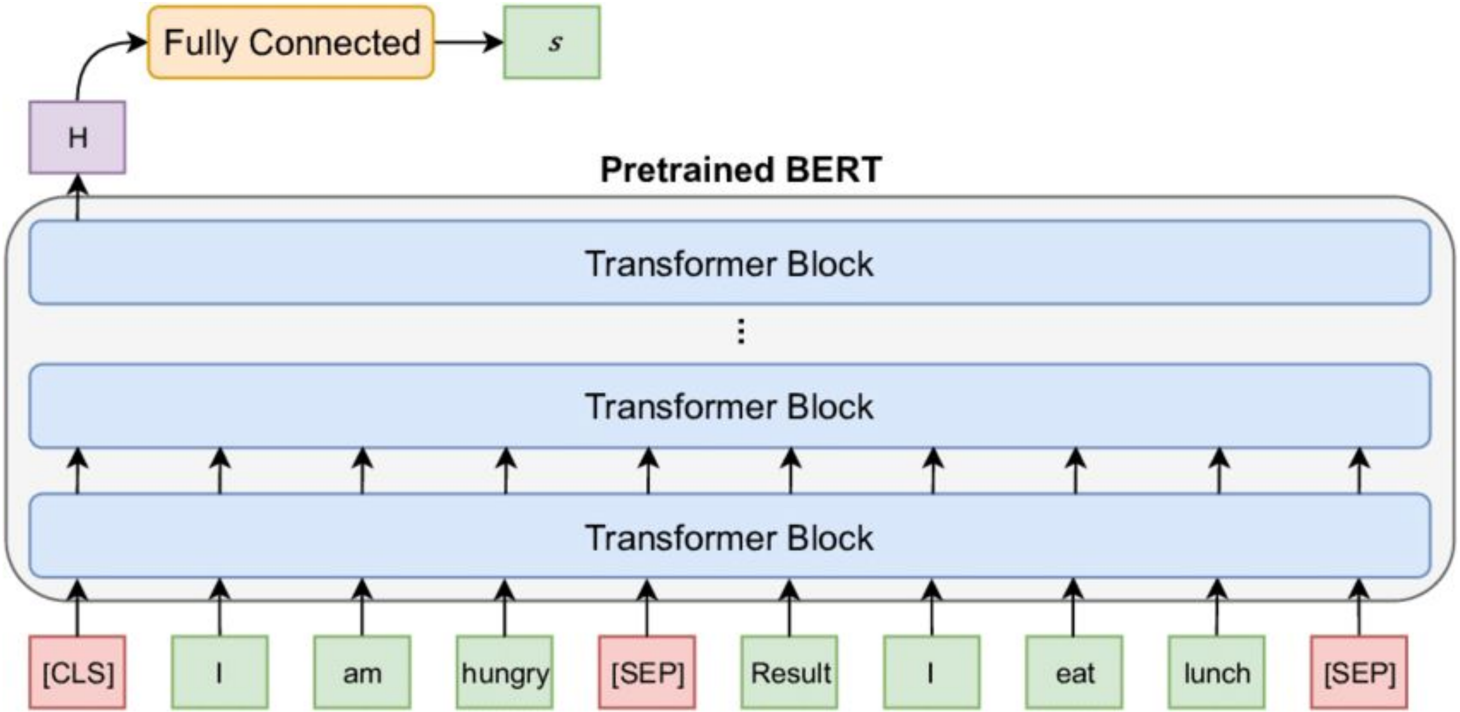


Fig 3. Bidirectional Encoder Representations from Transformers

Background Works

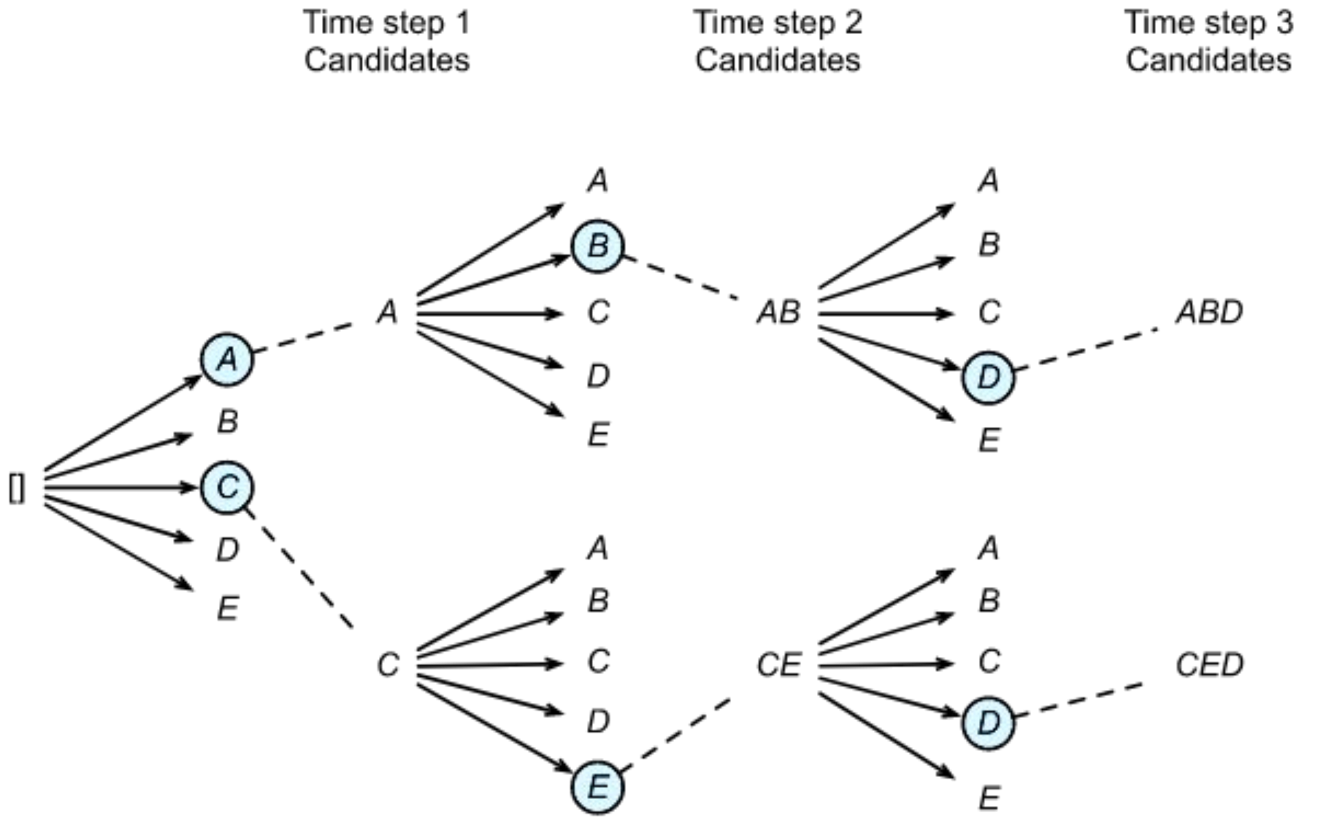


Fig 4. Beam Search

03

Related Works

Related Works

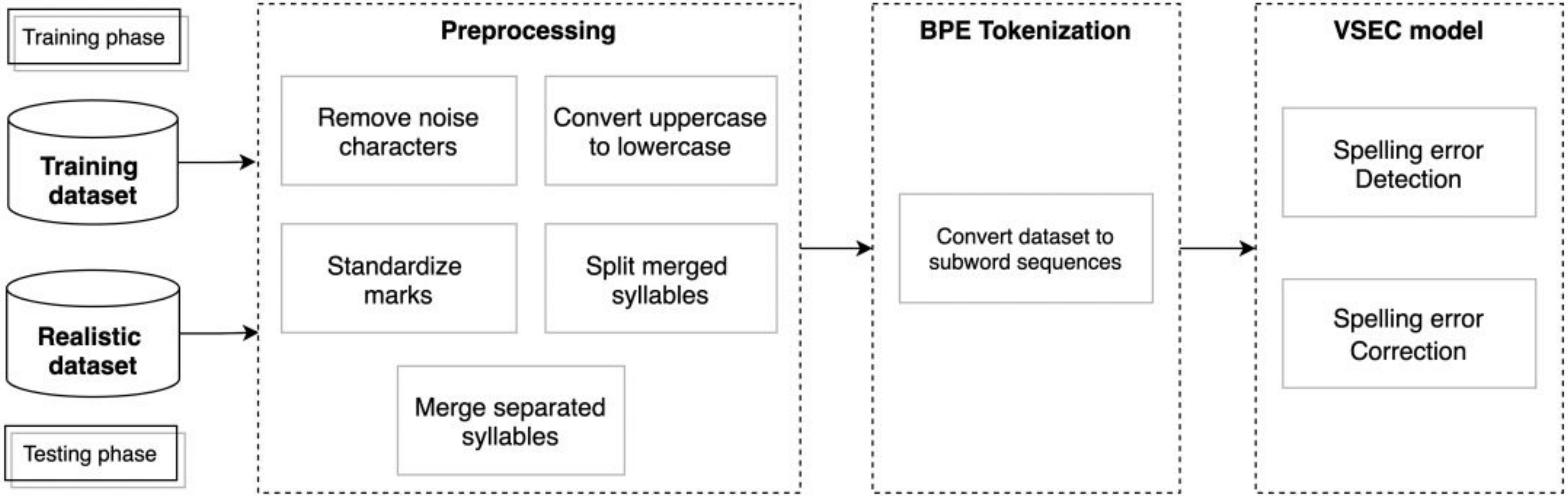


Fig 5. VSEC pipeline

Related Works

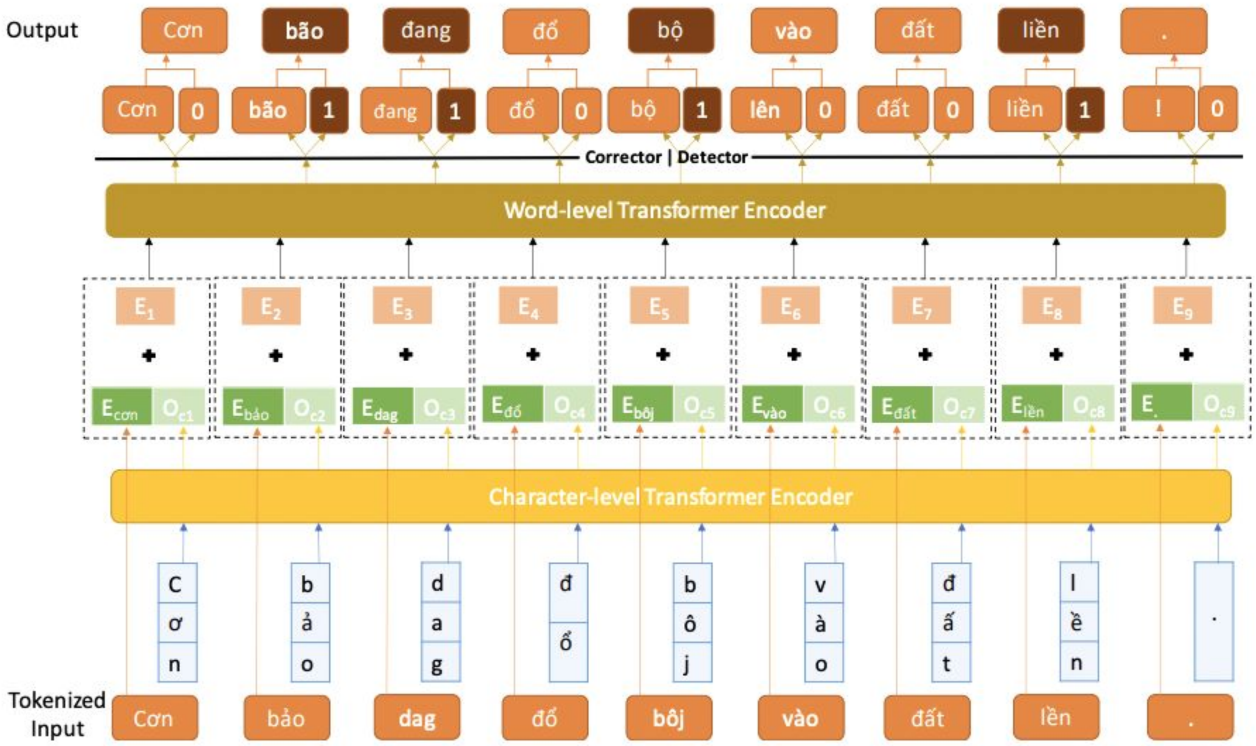


Fig 6. Hierarchical Transformer Encoders for Vietnamese Spelling Correction

Related Works

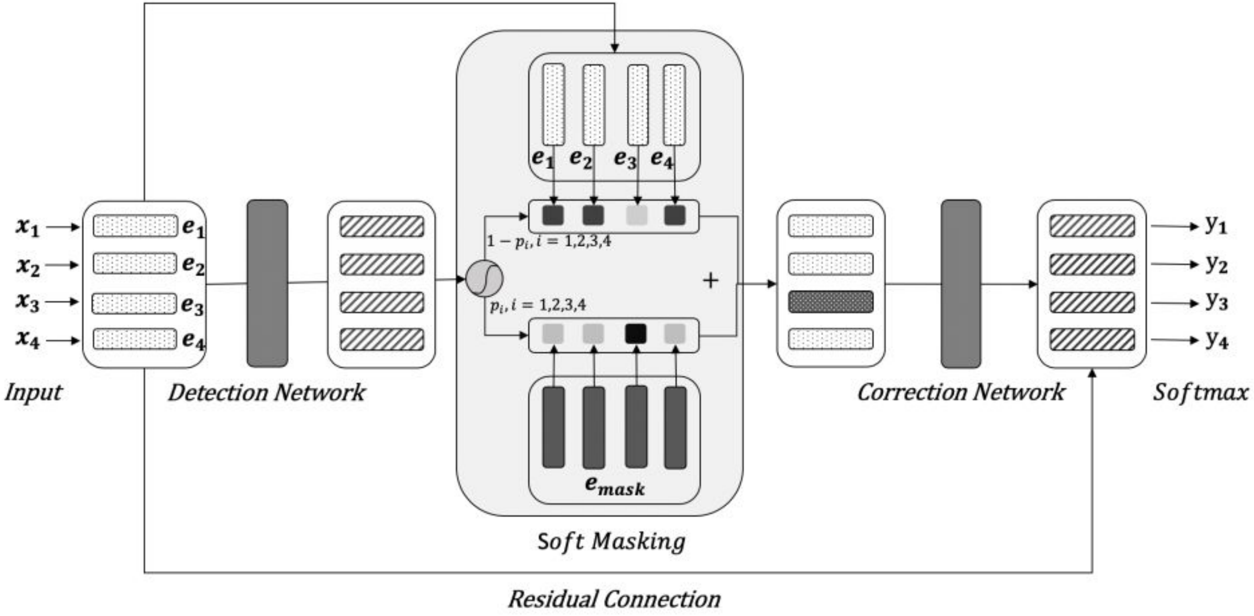


Fig 7. Soft-Masked BERT

04

Correction Model

Approach

- Leveraging a pre-trained Transformer Encoder
- Introduce a Tokenization Repair module that handle **re-tokenization of misspelled words**
    - “thôngthuofnwg” → “thông thuofnwg”
    - “v ăn vở” → “văn vở”

Tokenization Repair

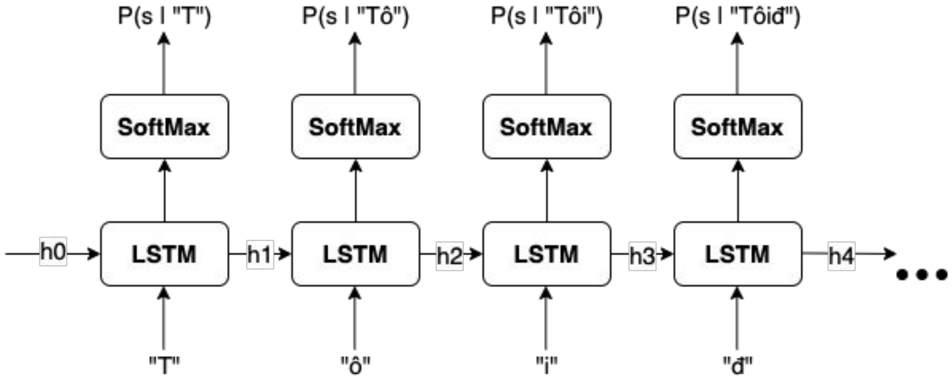


Fig 8. A unidirectional character language model

MODEL

- Character Language Model with LSTM

BEAM SEARCH

- Decode with Beam Search
- Without adding space

$$S_i = S_{i-1} - \log(\vec{p}(T_i \mid R_{i-1})) + P_{del}$$
$$R_i = R_{i-1}\ T_i$$

- Adding space

$$S_i = S_{i-1} - \log(\vec{p}(T_i \mid R_{i-1})) + P_{ins}$$
$$R_i = R_{i-1}\ T_i$$

- Penalize insert and delete

Tokenization Repair (Cont.)

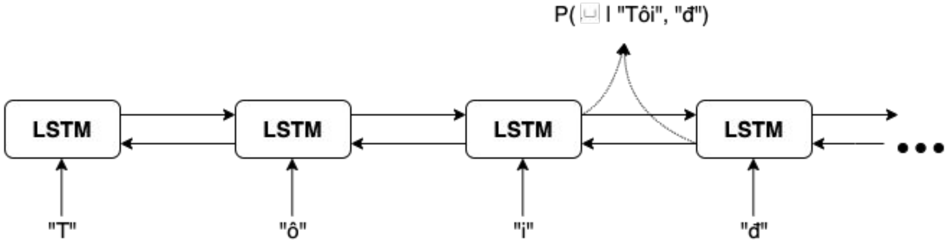


Fig 9. A Bidirectional Labeling Model

MODEL

- To enhance Unidirectional Language Model
- Bidirectional Labeling Model with BiLSTM

BEAM SEARCH

- Without adding space

$$S_i = S_{i-1} - \log(\vec{p}(T_i \mid R_{i-1}) * (1 - \overleftrightarrow{p})) + P_{del}$$
$$R_i = R_{i-1}\ T_i$$

- Adding space

$$S_i = S_{i-1} - \log(\vec{p}(T_i \mid R_{i-1}) * \overleftrightarrow{p}) + P_{ins}$$
$$R_i = R_{i-1}\ T_i$$

- Penalize insert and delete

Correction Model

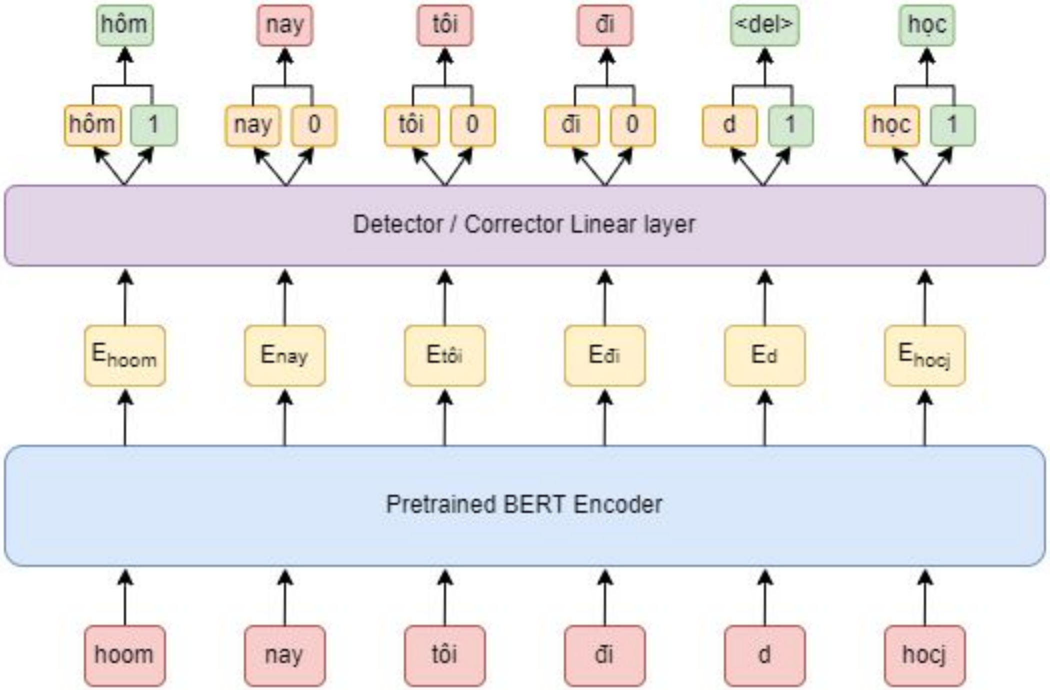


Fig 10. Correction Model

Correction Pipeline

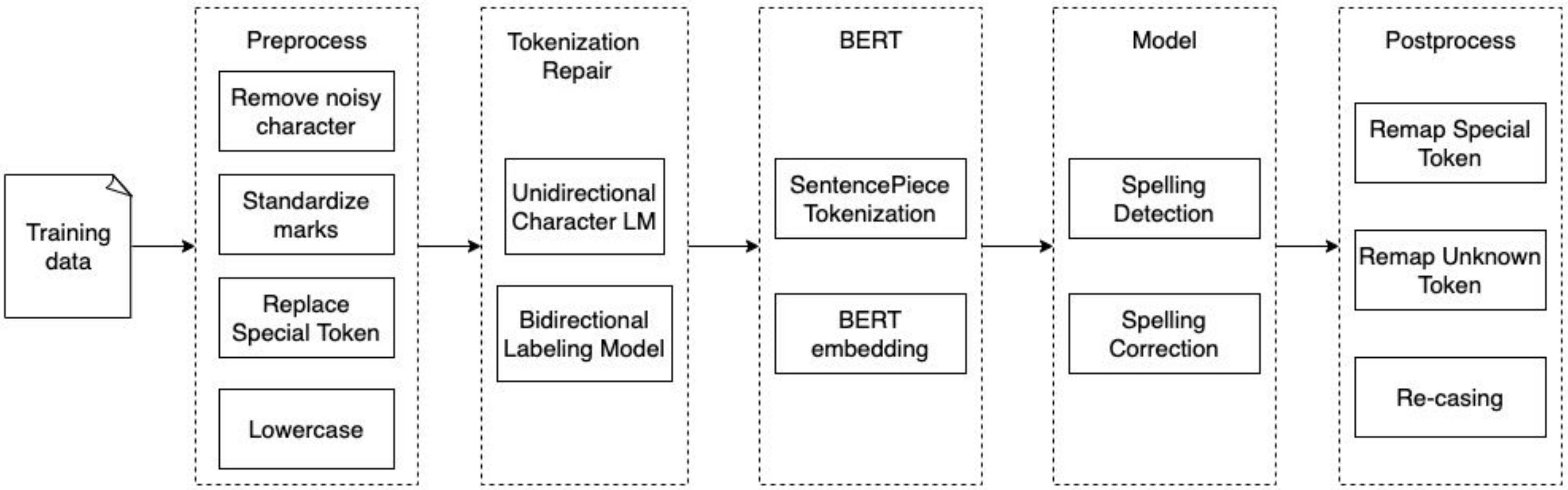


Fig 11. Contextual Spelling Correction Pipeline

05

Experiments & Result

Baselines

BASELINES

- Transformer model
  - Subword tokenization
  - Character tokenization
  - Word tokenization
- Hard-masked XLMR
- Soft-masked BERT
- BERT Fine-tuned

Baselines (cont)

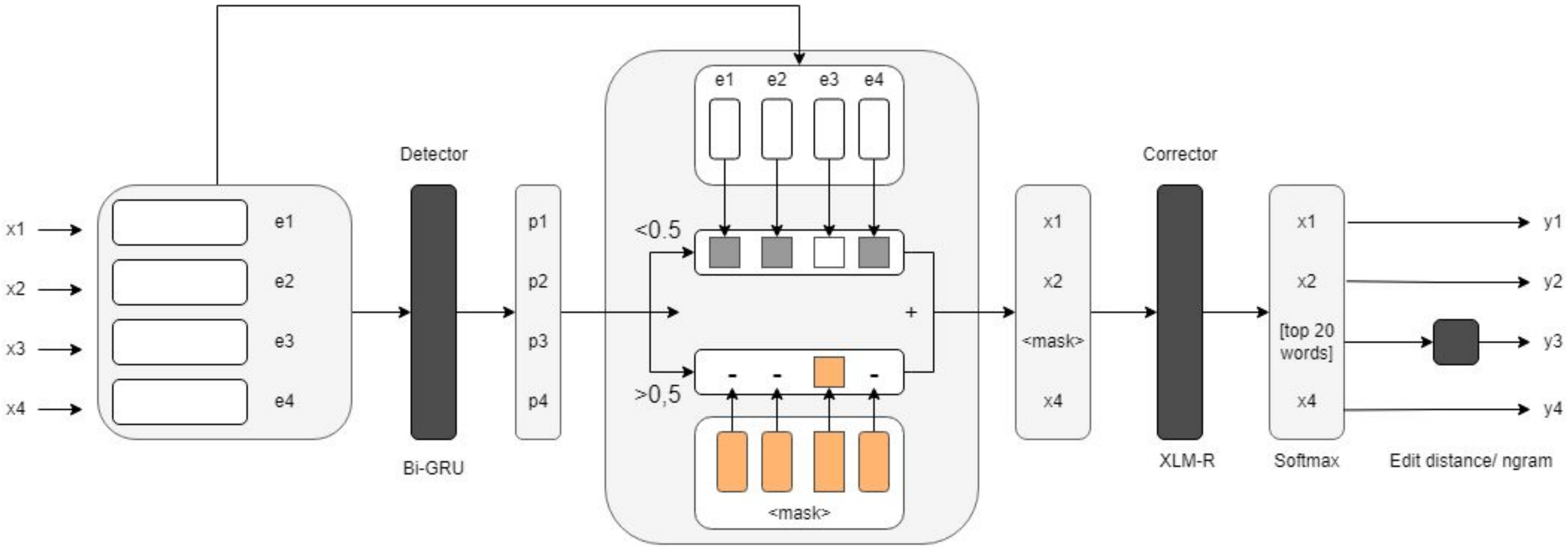


Fig 12. Hard-Masked XLMR

Baselines

BASELINES

- Transformer model
  - Subword tokenization
  - Character tokenization
  - Word tokenization
- Hard-masked XLMR
- Soft-masked BERT
- BERT Fine-tuned

EVALUATION METRICS

$$DetectionPrecision = \frac{TrueDetections}{TotalErrorDetected} \quad \text{(DP)}$$

$$DetectionRecall = \frac{TrueDetections}{TotalActualErrors} \quad \text{(DR)}$$

$$DetectionF1 - score = \frac{2*DP*DR}{DP+DR} \quad \text{(DF)}$$

$$CorrectionPrecision = \frac{TrueCorrections}{TotalErrorDetected} \quad \text{(CP)}$$

$$CorrectionRecall = \frac{TrueCorrections}{TotalActualError} \quad \text{(CR)}$$

$$CorrectionF1 - score = \frac{2*CP*CR}{CP+CR} \quad \text{(CF)}$$

Evaluation Metrics

Data Augmentation

PIPELINE

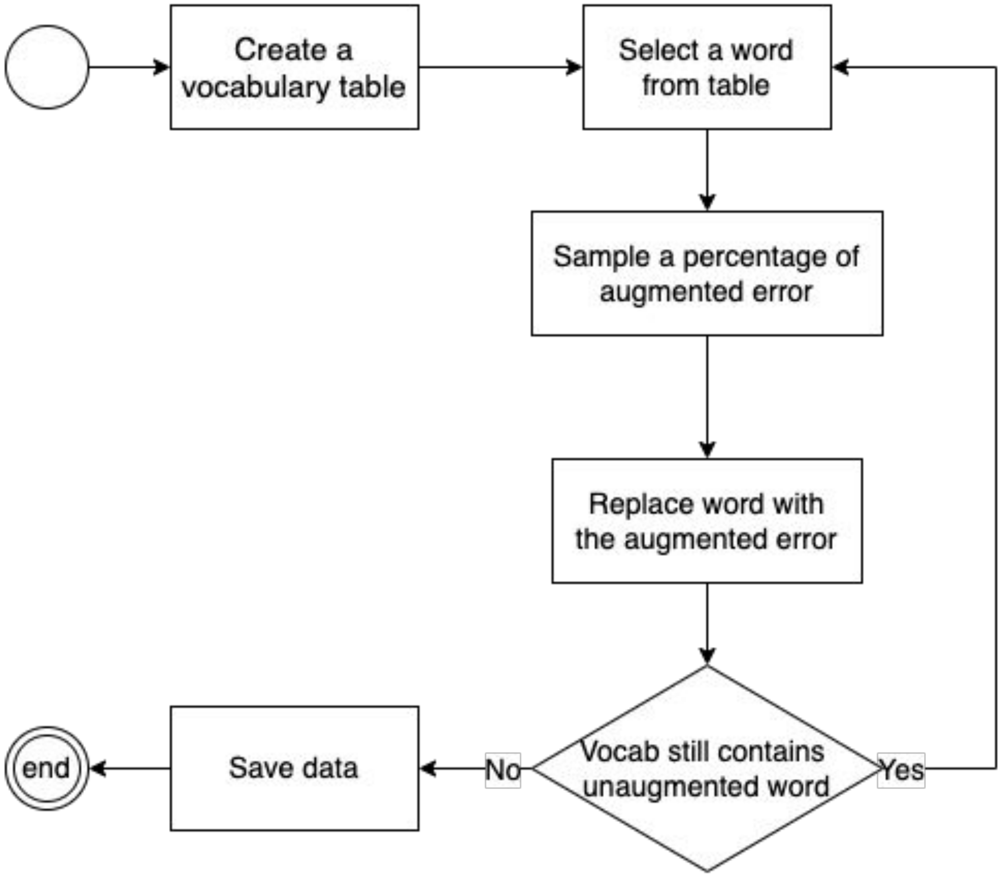


Fig 13:
Augmentation Pipeline

Data Augmentation (Cont.)

RULE BASED AUGMENTER EXAMPLES

```
telexAug.augment("tôi đi học",3)
```

```
['tôi đi hjoc', 'toio đi học', 'tooi đi học']
```

```
wsAug.augment("tôi đi học")
```

```
'tôiđi học'
```

```
keyAug.augment("tôi đi học", 1)
```

```
'tôi dsi học'
```

```
missDiaAug.augment("tôi đi học", 1)
```

```
'toi đi học'
```

```
wrDiaAug.augment("tôi đi học")
```

```
'tổi đi học'
```

```
spellBegAug.augment("tôi đi học rất vui")
```

```
'tôi đi học dất vui'
```

```
spellFinAug.augment("tôi đi học rất vui")
```

```
'tôy đi họt rất vui'
```

```
ed1Aug.augment("tôi đi học")
```

```
'tơi đi học'
```

Result

	Detection			Correction				
	DP	**DR**	**DF**	**CP**	**CR**	**CF**	**Acc_t**	**Acc_d**
Transformer + Character	96.22	73.24	83.17	71.73	54.50	62.0	71.73	74.54
Transformer + Word	71.89	84.23	77.57	40.0	46.86	43.16	40.0	55.63
Transformer + Subword	<u>98.84</u>	85.17	91.5	85.29	73.49	78.95	85.29	86.29
Hard-Masked XLMR	97.82	93.26	96.48	46.33	55.6	50.54	46.33	53.66
Soft-masked BERT	96.36	93.92	95.12	72.35	68.73	70.49	72.35	75.12
Our Model	97.20	<u>96.42</u>	<u>96.62</u>	<u>91.70</u>	<u>89.79</u>	<u>90.73</u>	<u>91.70</u>	<u>93.12</u>
- **Tokenization repair**	**98.36**	**97.31**	**97.83**	**92.07**	**91.09**	**91.58**	**92.07**	**93.61**

VSEC Result

	Detection			Correction				
	DP	**DR**	**DF**	**CP**	**CR**	**CF**	**Acc_t**	**Acc_d**
Transformer + Character	37.39	56.61	45.03	26.13	39.57	31.48	26.13	69.91
Transformer + Word	21.69	66.68	32.73	16.46	45.01	22.1	16.46	67.51
Transformer + Subword	82.0	76.95	79.39	71.82	67.4	69.56	71.82	87.59
Hard-Masked XLMR	85.03	78.11	82.24	53.91	42.44	47.49	53.91	68.78
Soft-masked BERT	81.55	79.2	80.35	59.8	59.76	58.24	59.8	78.43
Our Model	92.2	<u>85.23</u>	<u>88.58</u>	83.69	<u>77.36</u>	80.4	<u>83.69</u>	<u>90.77</u>
- **Tokenization repair**	**93.59**	**85.39**	**88.58**	<u>86.23</u>	**78.76**	**82.33**	**86.23**	**92.12**
VSEC (Paper, 2M)	89.1	76.9	82.6	82.6	71.3	76.5	-	-
VSEC (Paper, 5M)	<u>93.1</u>	81.3	86.8	**87.4**	76.3	<u>81.5</u>	-	-

VIWIKI Result

	Detection			Correction	
	DP	DR	DF	Acc_t	Acc_d
Transformer + Character	1.33	49.08	2.6	0.65	48.93
Transformer + Word	1.08	67.17	2.14	0.46	42.81
Transformer + Subword	4.38	58.64	8.16	3.36	76.73
Hard-Masked XLMR	23.84	60.67	34.22	8.96	48.32
Soft-masked BERT	21.72	58.76	31.72	12.38	60.54
Our Model	40.26	63.49	49.27	35.84	89.03
- **Tokenization repair**	<u>40.65</u>	<u>64.87</u>	<u>49.98</u>	<u>36.29</u>	<u>89.27</u>
Viwiki (Paper)	**66.96**	**70.92**	**68.88**	**64.29**	**96.01**

Effect of dataset size

	Detection			Correction	
	DP	**DR**	**DF**	**Acc_t**	**Acc_d**
500K	90.37	89.98	90.17	86.7	90.6
1M	92.98	92.4	92.69	87.13	91.24
2M	94.5	93.76	94.13	89.91	92.08
2.5M	**98.36**	**97.31**	**97.83**	**92.07**	**93.61**

VSEC Result

	Detection			Correction		
	DP	**DR**	**DF**	**CP**	**CR**	**CF**
Dictionary	s1	s1	s1	s1	s1	s1
Bi-LSTM	s1	s1	s1	s1	s1	s1
Transformer + Subword	s1	s1	s1	s1	s1	s1
Transformer + Character	s1	s1	s1	s1	s1	s1
Transformer + Word	s1	s1	s1	s1	s1	s1
Softmasked BERT	s1	s1	s1	s1	s1	s1
Our Model	s1	s1	s1	s1	s1	s1

VIWIKI Result

	Detection			Correction		
	DP	DR	DF	CP	CR	CF
Dictionary	s1	s1	s1	s1	s1	s1
Bi-LSTM	s1	s1	s1	s1	s1	s1
Transformer + Subword	s1	s1	s1	s1	s1	s1
Transformer + Character	s1	s1	s1	s1	s1	s1
Transformer + Word	s1	s1	s1	s1	s1	s1
Softmasked BERT	s1	s1	s1	s1	s1	s1
Our Model	s1	s1	s1	s1	s1	s1

Discussion

Discussion

Future Works

Demonstration App

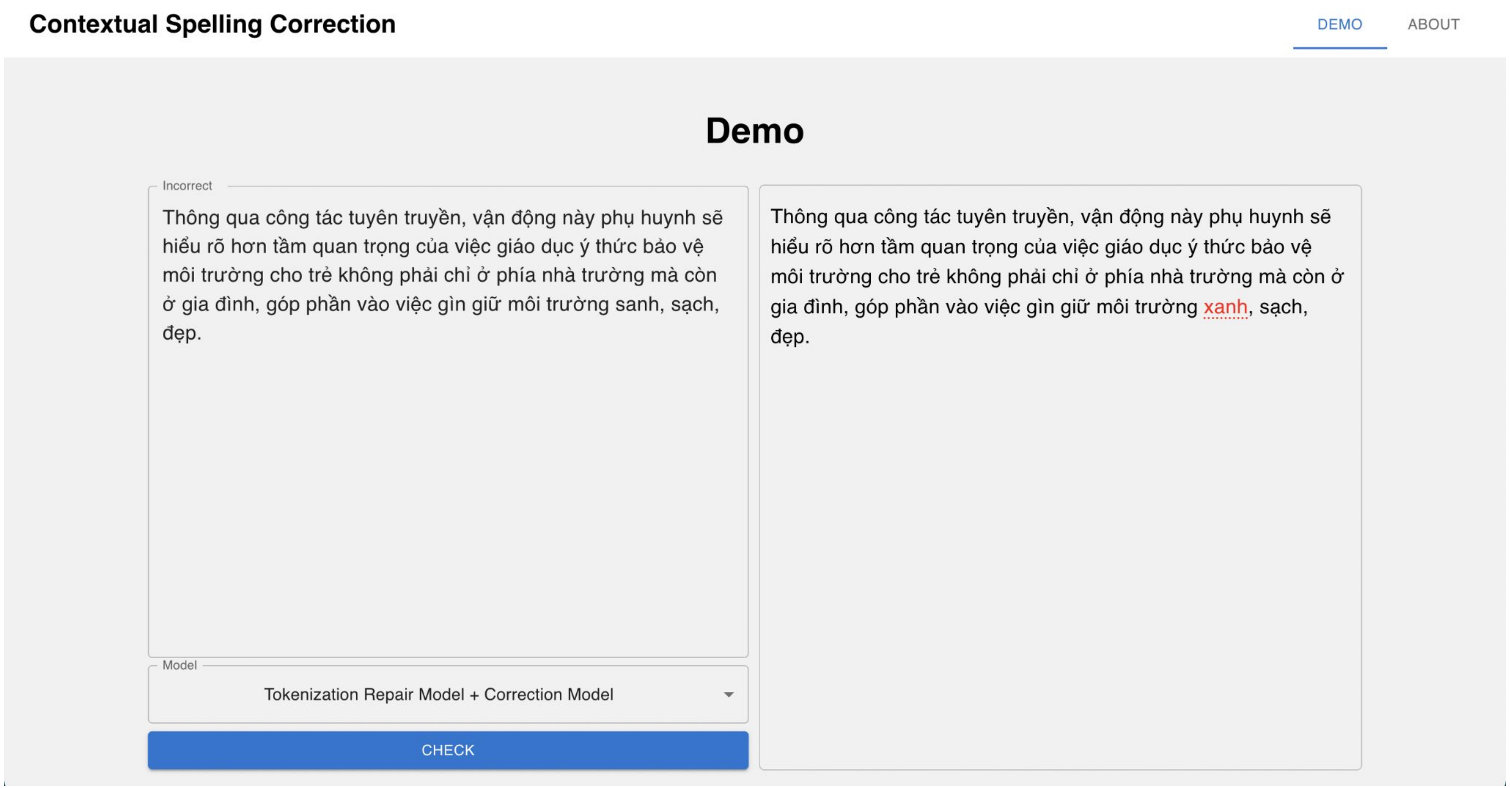


Fig 14. Demonstration Website

05

Resources