Vietnamese Tone Restoration

Group 4

Nguyễn Phú Trường, Lê Trần Hải Tùng, Lê Đức Tùng, Nguyễn Xuân Trường

Outline

- Problem & Data
- Methods
- Metric & Evaluation
- Demo

Problem & Data

Problem

Correct a sentence with no tone

- *“toi di da nang hom qua” → “tôi đi đà nẵng hôm qua”*

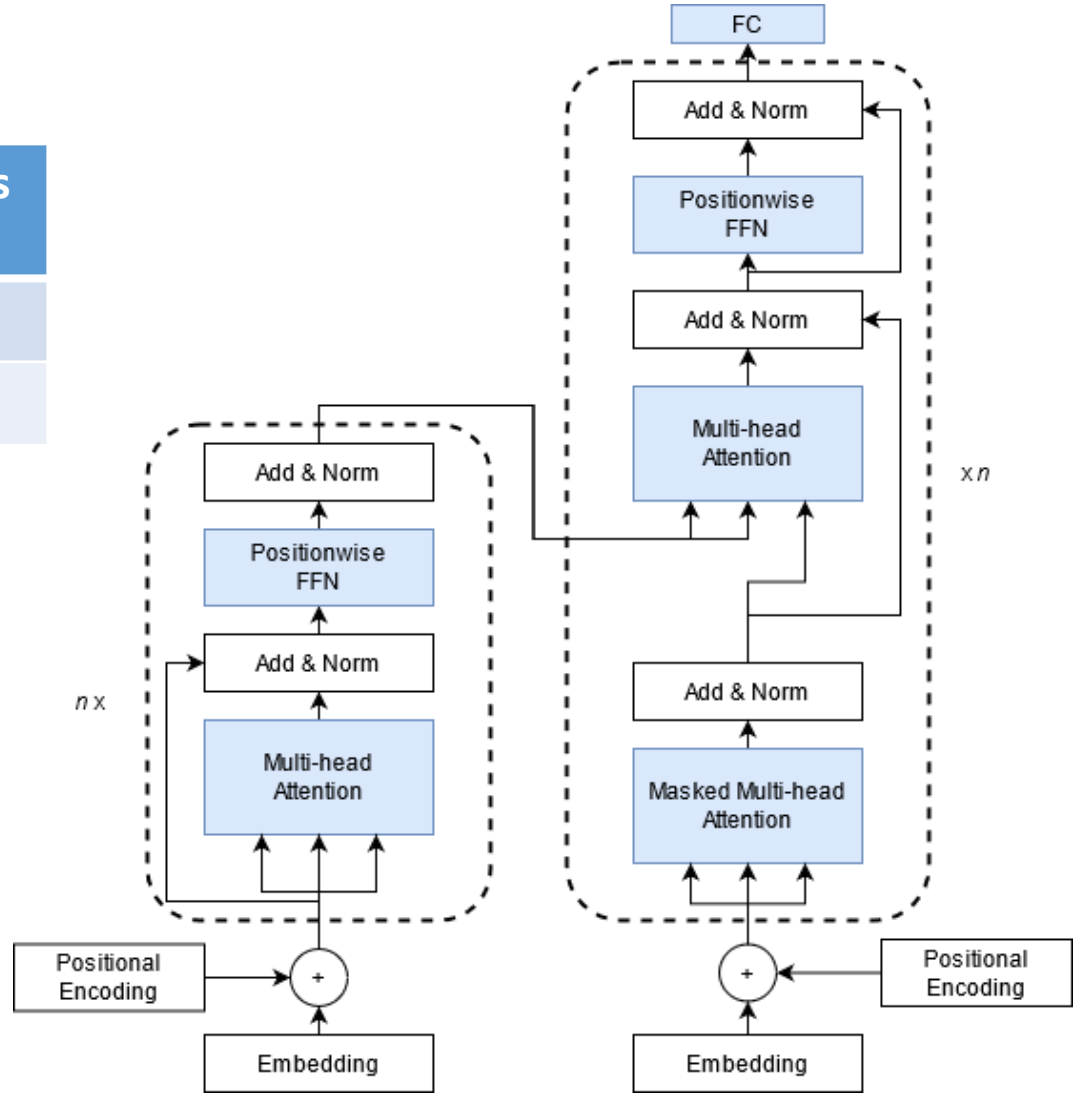
Data

- **Dataset 1**: includes *100K training sentence*, *1K testing sentence* from this repo. These are title from news article. We choose this dataset because they are short enough.
- **Dataset 2**: includes *200K training sentence*, *500 testing sentence* from our supervisor.

Methods

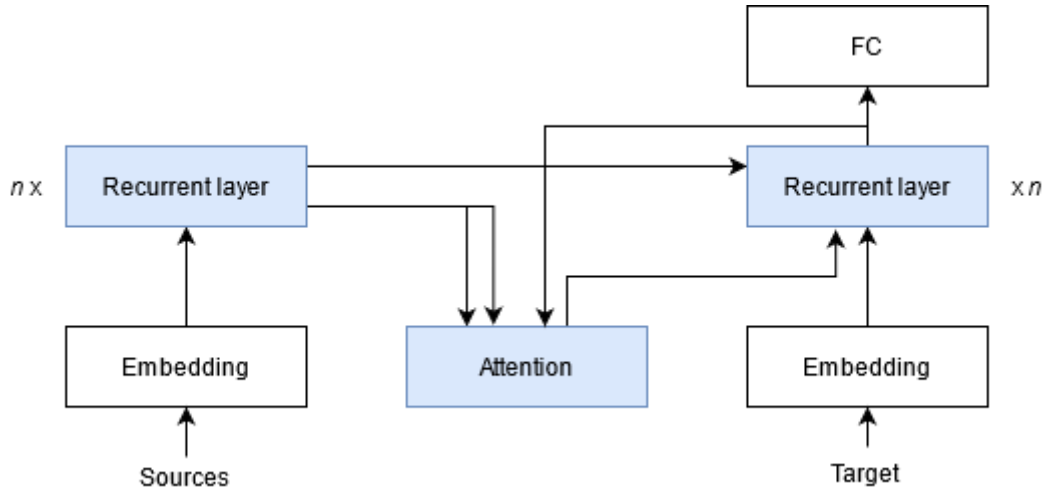
Transformers

Model	Hidden size	Num heads	Num layers	Learning rate	Epochs
SMALL	128	4	2	0.001	20
BASE	512	8	6	0.0001	30



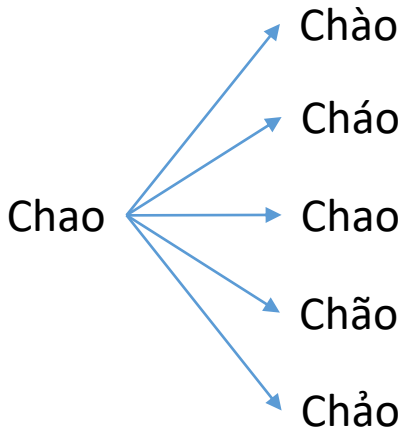
GRU Encoder Decoder

Model	Hidden size	Layers	Bi-directions	Learning rate	Epochs
SMALL	128	4	2	0.001	20



N-gram

- Tri-gram
- KnerseNey Smoothing



Evaluation

Metric

- **Mean accuracy** of all sentences
- *“toi di da nang hom qua”* → *“tôi đi đà năng hôm qua”*
- Ground truth: *“tôi đi đà nẵng hôm qua”*
- **Accuracy**: 5/6

Accuracy

Model	Accuracy on Dataset 1	Accuracy on Dataset 2
Transformers (SMALL)	**0.742**	0.766
GRU encoder decoder	0.712	0.661
N-gram	0.722	**0.813**

Model	Accuracy on Dataset 1	Accuracy on Dataset 2
Transformers (BASE)	**0.937**	**0.818** (One shot)

Training & evaluation time

Training

Model	Dataset 1 (h)	Dataset 2 (h)
Transformers (SMALL)	0.19	1.0
GRU encoder decoder	0.71	2.46
N-gram	*0.01*	*0.03*

Evaluation

Model	Dataset 1 (h)	Dataset 2 (h)
Transformers (SMALL)	0.01	0.005
GRU encoder decoder	0.007	0.004
N-gram	*2.27*	*2.23*

Thanks!

o(￣ヮ￣*)ゞ