Pre-training LongT5 for Vietnamese Multi-document Summarization Task

Nicolay Rusnachenko[1], The Anh Le[2], Ngoc Diep Nguyen[3]

rusnicolay@gmail.com, anhlt@vimaru.edu.vn, diepnn83@gmail.com

[1]Bauman Moscow State Technical University
[2]Vietnam Maritime University, Hai Phong, Viet Nam
[3]CyberIntellect, Moscow, Russia

Outline

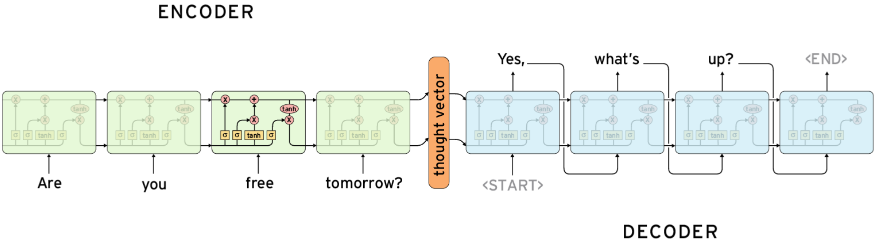
Structure of the talk:

- Overview of the Transformer Models
- Text Summarization in Vietnamese
- LongT5 Application

Transformer Models

Transofrmer Introduction

Before 2017:



- 2017 – Encoder-Decoder with Self-Attention (Transformer)[1]
  Language models Based on Transofrmers: BERT (language models)

[1] Ashish Vaswani et al. "Attention is All You Need". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 6000–6010. isbn: 9781510860964.

Long-Ranged Input for Transformers

Main limitation for input $X \in \mathbb{R}^N$:

- $O(N^2)$ original self-attention[1] computation complexity;

How to address this problem:

1. Sparse version[2] of self-attention: Reformer, Longformer[3]
2. #1 with Global Attention

[2] Joshua Ainslie et al. "ETC: Encoding long and structured inputs in transformers". In: *arXiv preprint arXiv:2004.08483* (2020).

[3] Iz Beltagy, Matthew E Peters, and Arman Cohan. "Longformer: The long-document transformer". In: *arXiv preprint arXiv:2004.05150* (2020).

Relative Position Encoding

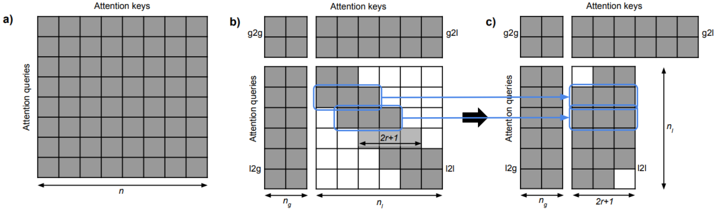
BERT[4] exploits absolute position encoding $X \in \mathbb{R}^N$.
ETC proposes **relative**:

- Now position is label $l_{i,j}$ of **connection** of $x_i \in X$ with other X
- Distance clipping: k – limit window
  - l_k outside after i,
  - l_{-k} outside radius k before i.
- **Result** in α_l^K – learnable vectors of relative positions

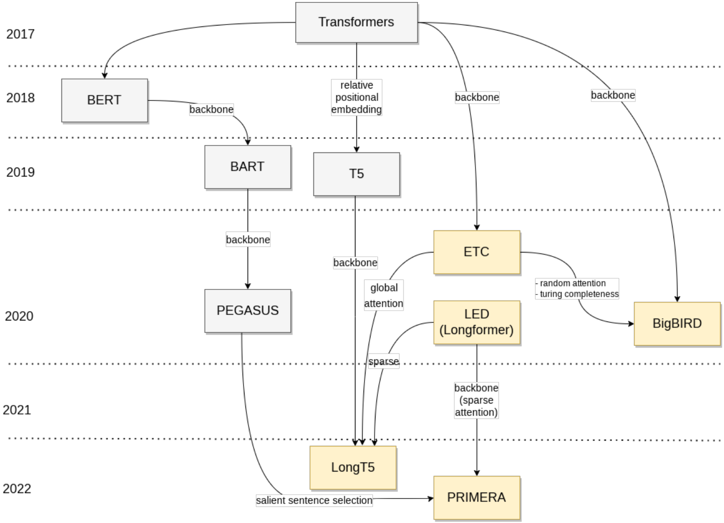
[4] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

Global + Local Attention

- n_l – main input components: **now windowed** (sparsed)
- n_g – global input components ($n_g << n_l$)

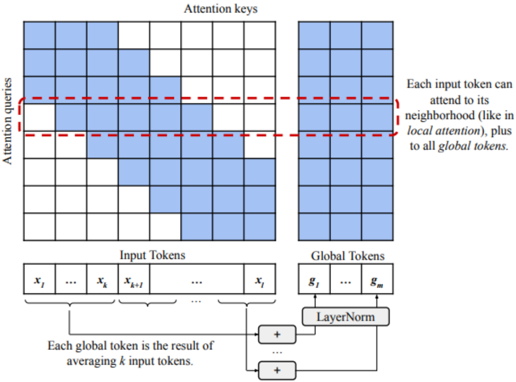


Affection on Future Models for Text Summarization



Model	Dataset	R-1	R-2	R-L
BART	XSum	45.14	22.27	37.25
PEGASUS	**CNN/DailyMail**	44.17	21.47	41.11
	Multi-News	47.52	18.72	24.91
	arXiv	44.21	16.95	38.83
T5	CNN/DailyMail	43.41	20.99	40.77
	Multi-News	47.48	18.60	24.31
	BigPatent	67.05	52.24	58.70
	arXiv	45.86	18.40	41.62
	PubMed	48.94	22.92	45.40
LED (16K)	arXiv	46.63	19.62	41.83
BigBird-PEGASUS	arXiv	46.63	19.02	41.77
	PubMed	46.32	20.65	42.33
	BigPatent	60.64	42.46	50.01
PRIMERA	arXiv	47.60	20.80	42.60
	Multi-News	49.90	21.10	25.90
LongT5 (4K)	CNN/DailyMail	42.49	20.51	40.18
	BigPatent	70.38	56.81	62.73
	arXiv	48.28	21.63	44.11
	PubMed	49.98	24.69	46.46

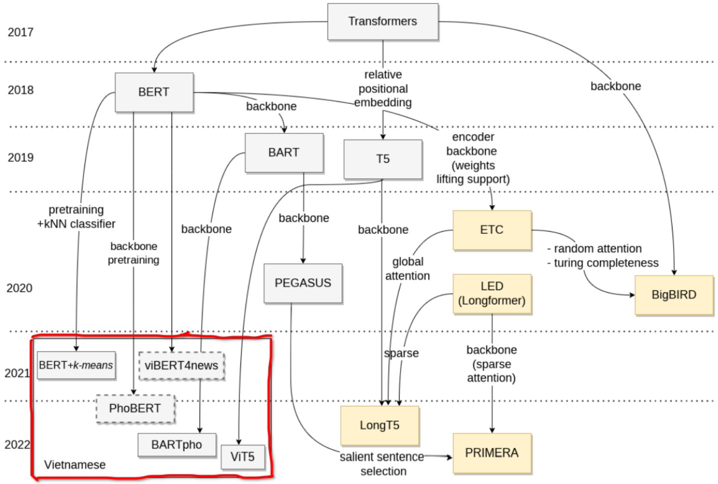
LongT5[5]



[5] Mandy Guo et al. "LongT5: Efficient Text-To-Text Transformer for Long Sequences". In: *Findings of the Association for Computational Linguistics: NAACL 2022*. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 724–736. doi: 10.18653/v1/2022.findings-naacl.55. url: https://aclanthology.org/2022.findings-naacl.55.

Vietnamese Models

Vietnamese models



Vietnamese Model Results

Model	ViMs R-1	ViMs R-2	VMDS R-1	VMDS R-2
LSA	62.5	36.0	62.9	37.0
LexRank	69.5	**46.4**	48.2	39.2
TextRank	62.8	41.6	66.2	40.8
SVR	64.5	39.7	66.9	44.3
SVMRank	63.5	41.0	67.4	46.2
MART	65.1	42.4	**70.2**	**49.6**
CNN	56.1	42.1	52.8	40.0
LSTM	**70.7**	43.1	52.5	39.6
XLM-R-large + *k-means*	—		77.4	51.2
PhoBERT-large + *k-means*	—		77.4	50.9
viBERT4news + *k-means*	—		**77.4**	**52.0**

Task

Multi-document Text Summarization[1]

VLSP2022 - Vietnamese abstractive multi-document summarization (Abmusu) task is to develop summarization systems that could create **abstractive summaries** automatically for a set of documents on a topic:

- **Domain:** Vietnamese news text.
- **Input:** Multiple news documents on the same topic.
- **Output:** Corresponding abstractive summary

1 https://vlsp.org.vn/vlsp2022

Collections[2]

Dataset	#doc	#samples	#docs per cluster	#words per document	#words per summary
NewsCorpus	14 896 998	–	–	–	–
VMDS	628	300	3.00	1308.00	153.00
ViMs	1 945	300	6.50	2208.00	192.00
VLSP2022$_{\text{train}}$	621	200	3.11	1925.75	168.48
VLSP2022$_{\text{valid}}$	304	100	3.04	1815.41	167.68
VLSP2022$_{\text{train+valid}}$	925	300	3.00	1853.00	162.00
VLSP2022$_{\text{test}}$	914	300	3.05	1762.40	153.05

2 https://paperswithcode.com/dataset/vnds

Training Setup

- LongT5$_{\text{LARGE}}$-TGlobal (2K/512) – is a case insensitive LongT5 version with Transient Global Attention mechanism with 2048/512 input/output tokens respectively, and size of the original T5$_{\text{LARGE}}$
- 10^6 docs from NewsCorpus processed by VnCoreNLP[3], syllables as '_'
- Consider paragraphs as subdocuments (`<doc-sep>`)
- **Tokenizer:** `SentencePiece`[4] model for vocabulary

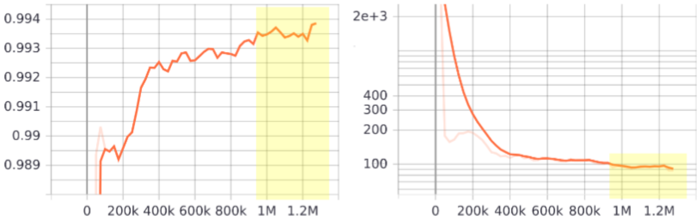
- **Input**: `<doc-sep>`, `<sent-sep>`, `<eos>` tokens
- **Summary:** PEGASUS principle sentence selection algo, 5 sentences.

3 https://github.com/vncorenlp/VnCoreNLP

4 https://github.com/google/sentencepiece

Pre-training Detail[7]

- Using `flaxformer`[5] library (GIN scripts over original `t5x`) and `jax`[6]
- 2 x NVidia A100 (40GB each), BatchSize=8, Speed: 4 steps/sec
- Train for `925K` steps (2.67 days) and then to `1.4M` (4.05 days)



5 https://github.com/google/flaxformer
6 https://github.com/google/jax
7 Yellow borders illustrates area we started to track the result behavior of summarization model

Fine-tunning

- **Joined:** ViMs, VMDS and VLSP2022$_{\text{train+valid}}$ datasets
- **Split:** 8:1:1 for train/dev/test

- ViLongT5$*$ (preliminary) 5K steps (BatchSize=8)
- ViLongT5 (final) 10K steps

Input	Result Summary
tại tuần_lễ đổi_mới sáng_tạo win-win của huawei , giám_đốc_điều_hành hội_đồng_quản_trị kiêm chủ_tịch hội_đồng quản_lý cơ_sở_hạ_tầng ict của huawei , ông david_wang , đã có bài phát_biểu quan_trọng mang tên ¨ổi_mới sáng_tạo và thắp sáng kỷ_nguyên 5 . 5g " . trong bài phát_biểu , ông wang nói nhiều về bước phát_triển tiếp_theo của công_nghệ 5g mà công_ty gọi là 5 . 5g , cũng như lộ_trình đổi_mới sáng_tạo của ngành 5-10 năm tới . " tầm nhìn đến năm 2025 , yêu_cầu dịch_vụ mạng đa_dạng và quy_mô lớn sẽ tạo ra tiềm_năng cực lớn cho thị_trường mới "	tại tuần_lễ đổi_mới sáng_tạo win-win của huawei , giám_đốc_điều_hành hội_đồng_quản_trị kiêm chủ_tịch hội_đồng quản_lý cơ_sở_hạ_tầng ict của huawei , ông david_wang , đã có bài phát_biểu quan_trọng mang tên ¨ổi_mới sáng_tạo và thắp sáng kỷ_nguyên 5 . 5g ⟨*eos*⟩ . trong bài phát_biểu , ông wang nói nhiều về bước phát_triển tiếp_theo của công_nghệ 5g mà công_ty gọi là 5 . 5g , cũng như lộ_trình đổi_mới sáng_tạo của ngành 5-10 năm tới . tầm nhìn đến năm 2025 , yêu_cầu dịch_vụ mạng đa_dạng và quy_mô lớn sẽ tạo ra tiềm_năng cực lớn cho thị_trường mới ⟨*eos*⟩
at huawei win-win innovation week, the executive director of the board and chairman of the ict infrastructure management board of huawei, mr. david wang, gave an important speech entitled "change" creative new and light up the 5. 5g era". during the keynote , mr. wang talked a lot about the next development of 5g technology which the company calls 5 . 5g , as well as the industry 's innovation roadmap for the next 5-10 years . "vision to 2025, diverse and large-scale network service requirements will create huge potential for new markets"	*at huawei win-win innovation week, the executive director of the board and chairman of the ict infrastructure management board of huawei, david wang, gave a keynote speech entitled " Innovate and light up the era of 5. 5g ⟨eos⟩. In his keynote , Mr. wang talked a lot about the next step of the development of 5g technology which the company calls 5. 5g, as well as the roadmap for the change innovation of the industry in the next 5-10 years with a vision to 2025, diverse and large-scale network service requirements will create huge potential for new markets ⟨eos⟩*

Experiment Results

Model	Rank	Dataset	Rouge Scores (F1)		
			R-1	R-2	R-L
$\text{hybrid}_{\text{the\_coach team}}$	#1	$\text{VLSP2022}_{\text{valid}}$	51.68	31.50	48.93
$\text{LexRank+MMR}_{\text{baseline}}$	#8	$\text{VLSP2022}_{\text{valid}}$	48.36	26.50	44.21
$\text{rule}_{\text{baseline}}$	#10	$\text{VLSP2022}_{\text{valid}}$	46.40	25.82	42.84
ViLongT5$*$	#13	$\text{VLSP2022}_{\text{valid}}$	45.70	24.83	42.85
$\text{anchor}_{\text{baseline}}$	#19	$\text{VLSP2022}_{\text{valid}}$	43.81	19.31	39.28
$\text{ViT5}_{\text{abstractive-baseline}}$	#20	$\text{VLSP2022}_{\text{valid}}$	31.29	30.77	27.97
$\text{hybrid}_{\text{the\_coach team}}$	#1	$\text{VLSP2022}_{\text{test}}$	49.62	29.37	47.01
$\text{LexRank+MMR}_{\text{baseline}}$	#6	$\text{VLSP2022}_{\text{test}}$	47.72	26.25	43.39
$\text{rule}_{\text{baseline}}$	#7	$\text{VLSP2022}_{\text{test}}$	46.27	26.11	42.73
ViLongT5	#10	$\text{VLSP2022}_{\text{test}}$	45.16	24.48	42.08
$\text{anchor}_{\text{baseline}}$	#19	$\text{VLSP2022}_{\text{test}}$	43.21	18.86	38.69
$\text{ViT5}_{\text{abstractive-baseline}}$	#20	$\text{VLSP2022}_{\text{test}}$	32.26	14.97	28.95

Conclusion

- We provide the survey of abstractive text summarization models
- We experiment with the finetuned version and result summaries represent the retelling of the most salient sentences
- The best submissions adopt hybrid summarization, which might be in further experiment as LexRank + ViLongT5

Thank you for attention!