LBMT Team at VLSP2022-Abmusu Hybrid Method with Text Correlation and Generative Models for Vietnamese Multi-Document Summarization

Tan-Minh Nguyen, Thai-Binh Nguyen, Hoang-Trung Nguyen, Hai-Long Nguyen,

Tam Doan Thanh, Ha-Thanh Nguyen, Thi-Hai-Yen Vuong

LBMT_Team@DSKTLab

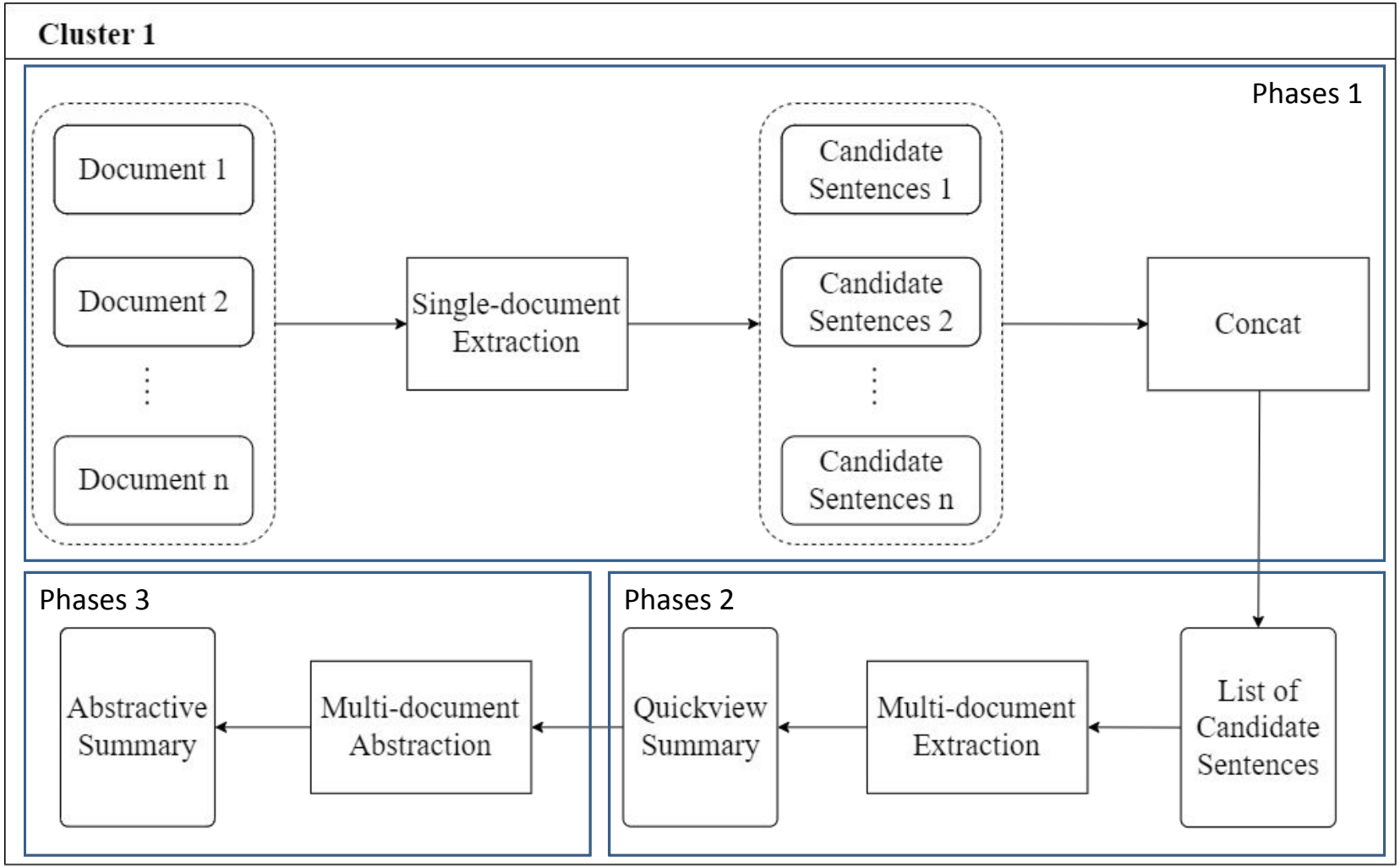
VNU - UET

Contents

Introduction

- Type of summarization task
  - Single/Multi-document summarization
  - Extractive/Abstractive summarization
- Difficulties
  - Cross-document relation
  - Larger searching space
  - Redundancy, duplication information
  - Number of documents in a cluster is not fixed
  - Diverse document lengths

Overall Architecture



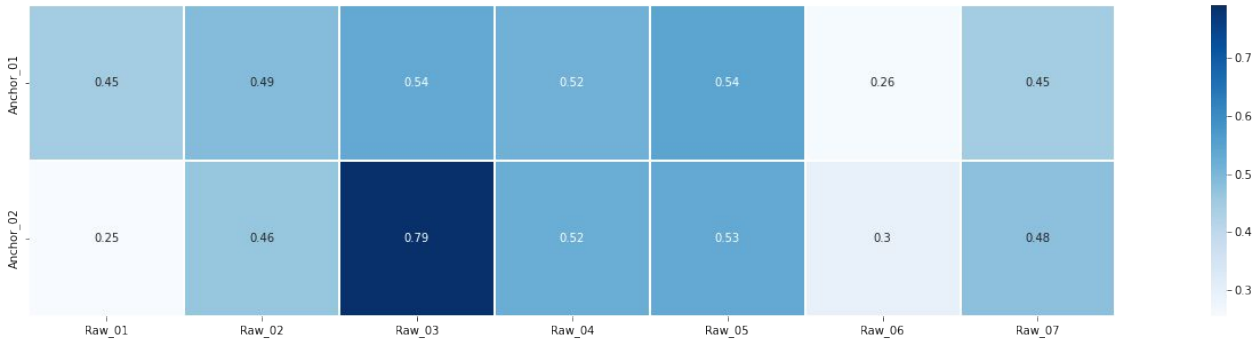
Phase 1: Single document extraction

- Calculate text correlation between anchor text and body text
  - Anchor text: $A = \{a_i | i \in [1, n]\}$ where n is the number of sentences
  - Body text: $\mathrm{B} = \{b_j | j \in [1, m]\}$ where m is the number of sentences
- Measure text correlation using Cosine similarity $S_c(A, b_j)$

$$S_c(A, b_j) = \sum_{i=1}^{n} S_c(a_i, b_j)$$

- Ranking candidate sentences by its similarity score

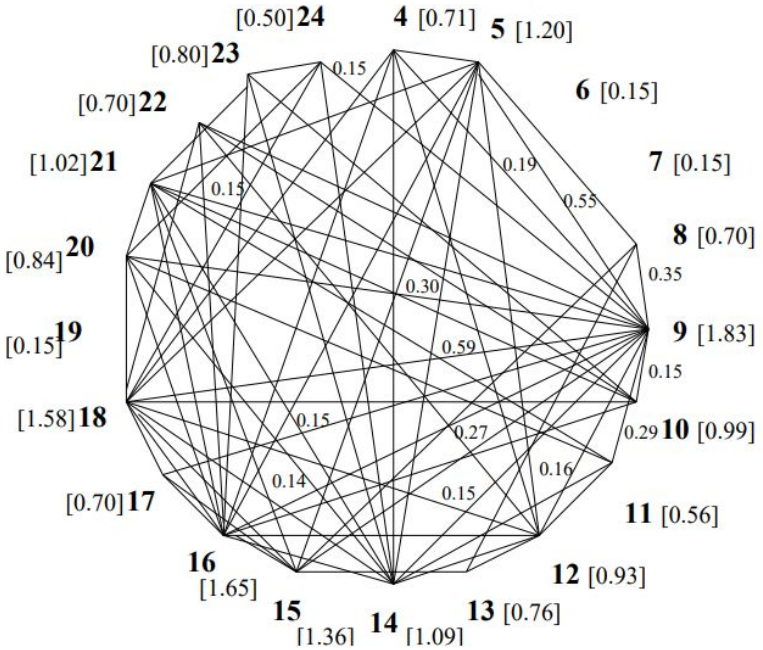
Phase 1: Single document extraction



Anchor: "Mua xe mới bị đại lý bắt chẹt 'bán bia kèm lạc', lại chờ quá lâu, nhiều khách hàng đã chuyển hướng sang mua ô tô cũ, chạy lướt, chấp nhận giá cao. Đặc biệt, các mẫu xe SUV 'hot' giá cao hơn cả xe mới, dân buôn được mùa lãi đậm."	**Anchor: "Buying a new car was caught by the dealer "selling beer with peanuts", waiting too long, many customers turned to buying used cars, surfing, accepting high prices. In particular, the 'hot' SUV models are priced higher than new cars, and traders get a big payoff season."**
Raw_03: "Cũng vì thế, nhiều tháng nay, thị trường ô tô cũ được dịp "lên ngôi," đặc biệt một số dòng xe SUV hot đang có giá bán cao hơn cả giá bán lẻ đề xuất của xe mới."	Raw_03: "Because of that, for many months now, the used car market has had the opportunity to "take the throne," especially some hot SUVs that are selling for more than the suggested retail price of new cars."
Raw_06: "Hiện tại tôi đã mua được chiếc Tucson 2021 dung tích 2.0L bản xăng đặc biệt chạy lướt 1.300km với giá 1,23 tỷ đồng."	Raw_06: "Currently, I have bought a 2.0L Tucson 2021 with a special gasoline version that runs for 1,300km for 1.23 billion VND."

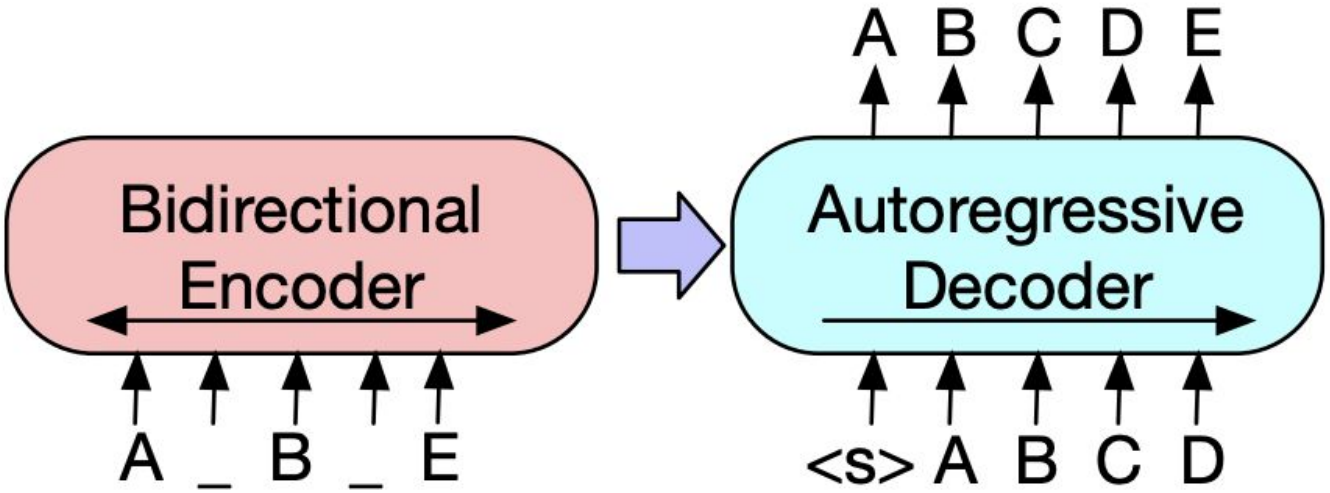
Phase 2: Multi-document extraction

- Construct TextRank graph for each cluster to ranking importance of sentences
  - A vertex is a sentence from previous phase
  - Choose top candidate sentences
- Filter high correlation sentences among list of candidates
- Generate a quickview summarization of this cluster



Phase 3: Multi-document abstraction

- Utilize the pre-trained model BARTpho, ViT5
- Fine-tune these models with (quickview, label) pairs dataset
- Produce abstractive summaries

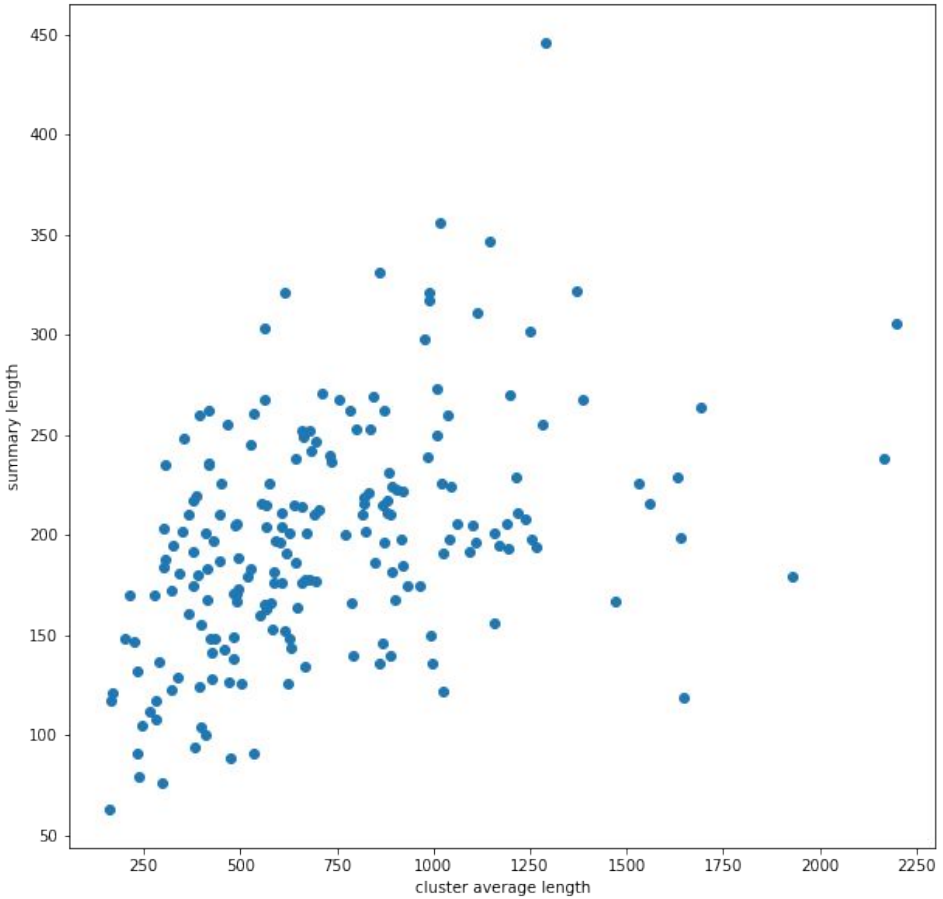


Data analysis

	Train	Valid	Test
# clusters	200	100	300
# documents/cluster	2-5	2-5	2-5
# avg document	3.105	3.04	3.05
Max document len	3474	4619	4291
Min document len	148	132	73
Avg document len	715.23	701.65	667.13
Max summary len	446	380	
Min summary len	63	83	
Avg summary len	197.74	202.35	

Cluster-summary lengths relation

- A cluster and summary lengths have a linear relation
- The ratio between a cluster and summary is mainly 2.0 to 4.0



Experiments and Results

PUBLIC TEST RESULTS

	R2-F1	R2-P	R2-R	R1-F1	R1-P	R1-R	RL-F1	RL-P	RL-R
Single document extraction	0.2536	0.2094	0.3465	0.4902	0.4267	0.6037	0.4550	0.3966	0.5592
Multi-document extraction	**0.3149**	0.2566	**0.4577**	**0.5178**	0.4408	**0.6696**	**0.4881**	0.4160	**0.6305**
BARTpho-base	0.2607	**0.2612**	0.2911	0.4788	**0.4870**	0.5041	0.4428	**0.4510**	0.4658

Experiments and Results

PUBLIC LEADERBOARD

User	R2-F1	R2-P	R2-R	R1-F1	R1-P	R1-R	RL-F1	RL-P	RL-R
thecoach_team	**0.3150**	0.2492	**0.4652**	0.5168	0.4307	**0.6804**	**0.4893**	0.4080	**0.6436**
minhnt2709	0.3149	0.2566	0.4577	**0.5178**	0.4408	0.6696	0.4881	0.4160	0.6305
TheFinalYear	0.2931	0.2424	0.4137	0.5080	0.4372	0.6433	0.4733	0.4081	0.5984
TungHT	0.2875	**0.2978**	0.2989	0.4902	**0.5061**	0.5012	0.4536	**0.4681**	0.4638
ngtiendong	0.2841	0.2908	0.2949	0.4932	0.5024	0.5035	0.4547	0.4634	0.4642

Example

Phase 1	Phase 2	Phase 3
Trong khi đó, sàn giao dịch tiền mã hóa Coinbase cũng mất hơn 1,1 tỷ USD trong cùng khoảng thời gian, chủ yếu đến từ việc các đồng coin mất giá liên tiếp. Đồng thời, việc các công ty khai thác lớn chốt lãi những khoản vay là một điều đáng bận tâm. Song, Bitcoin vẫn mang lại lợi nhuận cho các hoạt động thương mại quy mô lớn, với mỗi khối được khai thác trị giá khoảng 6,25 BTC, hoặc khoảng 120.000 USD theo giá hiện nay. Đáng nói, bên mua chiếm tới 90%, đi ngược với xu hướng bán tháo trên thị trường tiền mã hóa toàn cầu. "Khi thị trường liên tục biến động, các nhà đầu tư muốn tìm kiếm những bến đỗ an toàn để giao dịch và lưu trữ tài sản", ông Lionel Lim - CEO tại DBS Digital Exchange - bình luận.	Trong khi đó, sàn giao dịch tiền mã hóa Coinbase cũng mất hơn 1,1 tỷ USD trong cùng khoảng thời gian, chủ yếu đến từ việc các đồng coin mất giá liên tiếp. Theo đánh giá từ tác giả Ian Evenden của Tom’s Hardware, giá tiền mã hóa giảm mạnh khi các thợ đào bán bớt tài sản của họ để trang trải chi phí và mở rộng dòng vốn. Song, Bitcoin vẫn mang lại lợi nhuận cho các hoạt động thương mại quy mô lớn, với mỗi khối được khai thác trị giá khoảng 6,25 BTC, hoặc khoảng 120.000 USD theo giá hiện nay.	Giá tiền mã hoá hôm nay 1/9 : Giảm nhẹ trên các sàn giao dịch thế giới . Trong 24 giờ qua , giá Bitcoin giao dịch ở mức hơn 20.000 USD , giảm khoảng 1% trong 24 giờ qua . Ông Lionel Lim - CEO tại DBS Digital Exchange - bình luận rằng dự báo này là " khó xảy ra " . Nhu cầu đầu tư vào những tài sản kỹ thuật số của các tổ chức đầu tư và văn phòng gia đình đang tăng mạnh khi các thợ đào bán bớt tài sản của họ để trang trải chi phí và mở rộng dòng vốn .

Conclusion

- We propose a three-phase pipeline for the Abmusu task.
- Both extractive and abstractive approaches were experimented.
- The proposed method achieves competitive results in the VLSP 2022 competition.

Future Work

- In the future, we will consider the relationship between sentences and documents to improve the quality of summaries.
- Expand research in Vietnamese grammar to enhance summary performance

Thanks for watching

LBMT_Team@DSKTLab