BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC CÔNG NGHỆ TP.HCM

LUẬN VĂN THẠC SĨ
CHUYÊN NGÀNH CÔNG NGHỆ THÔNG TIN

ÁP DỤNG MÔ HÌNH ĐỒ THỊ VÀO BÀI TOÁN TÓM TẮT VĂN BẢN TIẾNG VIỆT

BẠCH QUỐC HUY
GVHD:TS. LÊ THỊ NGỌC THƠ

TPHCM, Tháng 10/2018

Báo cáo luận văn Thạc Sĩ

Nội dung

1. Lý do chọn đề tài
2. Nội dung nghiên cứu
3. Đánh giá hiệu quả của đề tài
4. Kết luận

Lý do chọn đề tài

- ❖ Sự phát triển mạnh mẽ của mạng internet đem đến cho chúng ta một kho thông tin khổng lồ để tìm hiểu và khai thác.
- ❖ Bài toán được đặt ra:
- ➢ Làm thể nào có thể tìm kiếm và khai thác một cách hiệu quả lượng thông tin khổng lồ trên?

Lý do chọn đề tài

- ❖ Những vấn đề cần được giải quyết:
  - ➢ Đâu là những thông tin thuộc lĩnh vực chúng ta quan tâm?
  - ➢ Những thông tin, dữ liệu này có ích đối với chúng ta hay không?
- ❖ Phương án giải quyết:
  - ➢ Cần một công cụ tóm tắt phần nào nội dung các luồng thông tin.
  - ➢ Cung cấp thông tin khái quát cho người đọc

Lý do chọn đề tài

Các hệ thống tóm tắt văn bản đã được đề xuất:

- LexRank
- Chức năng tự động tóm tắt văn bản trong Microsoft Word
- Hệ thống tóm tắt online Text Compactor

Lý do chọn đề tài

Hiệu quả của các hệ thống tóm tắt văn bản tự động trên khi áp dụng vào tiếng việt là chưa cao do:

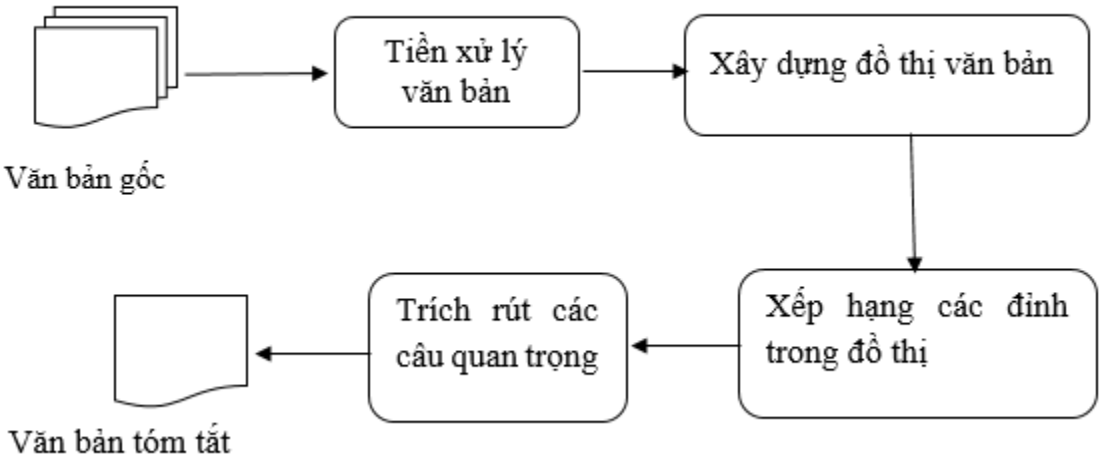
- Được xây dựng phục vụ cho ngôn ngữ tiếng Anh.
- Cấu trúc của tiếng Việt rất phức tạp.

Vấn đề cần đưa ra các ý tưởng để xây dựng nên một hệ thống tóm tắt văn bản tiếng Việt có hiệu quả.

Nội dung nghiên cứu

- Áp dụng mô hình đồ thị vào bài toán tóm tắt văn bản Tiếng Việt
- Đánh giá hiệu quả của các hệ thống tóm tắt văn bản tiếng việt xây dựng trên mô hình đồ thị thông qua các phương pháp tính toán độ tương đồng giữa 2 câu:
  - Độ đo cosine
  - Độ đo Euclidean
  - Khoảng cách theo thuật toán TextRank gốc
  - Số từ đơn giống nhau
  - Số cụm từ giống nhau

Nội dung nghiên cứu



Mô hình tóm tắt văn bản sử dụng phương pháp đồ thị

Nội dung nghiên cứu

- ❖ Các bước thực hiện
  - ➢ Tiền xử lý văn bản
  - ➢ Xây dựng đồ thị văn bản
  - ➢ Xếp hạng các đỉnh của đồ thị
  - ➢ Trích xuất các câu theo thứ tự từ cao xuống thấp
  - ➢ Hoàn thành văn bản tóm tắt

Nội dung nghiên cứu

❖ Phương pháp tính độ tương tự giữa các câu:

➢ Sử dụng độ đo cosine

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

➢ Sau khi các câu đã được biểu diễn dưới dạng vecto bằng cách sử dụng Doc2vec chúng ta sẽ đi tính toán độ tương tự của chúng dựa vào công thức trên

Nội dung nghiên cứu

❖ Phương pháp tính độ tương tự giữa các câu (tiếp):

➢ Sử dụng độ đo Euclidean

Cho 2 vecto p = (p1, p2,..., pn) and q = (q1, q2,..., qn)

Khoảng cách giữa p và q

$$d(\mathbf{p},\mathbf{q}) = d(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

➢ Sau khi các câu đã được chuyển đổi bằng vecto bằng cách sử dụng Doc2vec chúng ta sẽ đi tính toán độ tương tự của chúng dựa vào công thức trên

Nội dung nghiên cứu

❖ Phương pháp tính độ tương tự giữa các câu (tiếp)

➢ Độ tương tự giữa 2 câu được tính theo thuật toán TextRank gốc

$$\text{Similarity}(Si, Sj) = \frac{|\{Wk | Wk \in Si \wedge Wk \in Sj\}|}{\log(|Si|)+\log(|Sj|)}$$

➢ Trong đó

- Similarity(S_i, S_j): độ tượng tự giữa 2 câu
- W_k : Số từ giống nhau giữa 2 câu
- Log($|S_i|$) + Log($|S_j|$): Hệ số chuẩn hóa

Nội dung nghiên cứu

❖ Phương pháp tính độ tương tự giữa các câu (tiếp)

➢ Độ tương tự giữa 2 câu được tính theo số từ đơn giống nhau

$$Similarity(S_i, S_j) = |\{Wk | Wk \in Si \wedge Wk \in Sj\}|$$

➢ Trong đó:

- Similarity(S_i, S_j): Độ tương tự giữa 2 câu.
- Wk: Số từ đơn giống nhau giữa hai câu.

Nội dung nghiên cứu

- ❖ Phương pháp tính độ tương tự giữa các câu (tiếp)
  - ➢ Độ tương tự giữa 2 câu được tính theo số từ đơn giống nhau
  - ➢ Ví dụ:

S1: Tôi đi học.

S2: Tôi đi ăn cơm.

Similarity (S1, S2) =2

Nội dung nghiên cứu

❖ Phương pháp tính độ tương tự giữa các câu (tiếp)

➢ Độ tương tự giữa 2 câu được tính theo số từ cụm từ tiếng việt có nghĩa giống nhau:

$$Similarity(S_i, S_j) = |\{Wk | Wk \in Si \ ^\wedge Wk \in Sj\}|$$

➢ Trong đó:

- Similarity(S_i, S_j): Độ tương tự giữa 2 câu.
- Wk: Số cụm từ giống nhau.

Nội dung nghiên cứu

- ❖ Phương pháp tính độ tương tự giữa các câu (tiếp)
- ➢ Độ tương tự giữa 2 câu được tính theo số từ cụm từ tiếng việt có nghĩa giống nhau ví dụ:
- S1: Tuy_nhiên, trên thực_tế, đột_qụy thiếu máu não là dạng đột_quỵ thường gặp hơn.
- S2: Cứ 3 bệnh_nhân đột_qụy sẽ có 2 người đột_quỵ thiếu máu não.
- Similarity (S1, S2) =5

Nội dung nghiên cứu

❖ Phương pháp đánh giá

➢ Sử dụng phương pháp ROUGE với 3 đại lượng Recall, Precision, F-score :

$$R = \frac{Tổng\ số\ từ\ giống\ nhau\ giữa\ tóm\ tắt\ tự\ động\ và\ tóm\ tắt\ của\ con\ người}{Số\ từ\ văn\ bản\ tóm\ tắt\ của\ con\ người}$$

➢ Recall: là tỉ lệ các tài liệu có liên quan đến tài liệu truy vấn và trên thực tế được trích xuất trả về chính bằng ROUGE với n=1.

Nội dung nghiên cứu

❖ Phương pháp đánh giá (tiếp)

$$P = \frac{Tổng\ số\ từ\ giống\ nhau\ giữa\ tóm\ tắt\ tự\ động\ và\ tóm\ tắt\ của\ con\ người}{Số\ từ\ của\ hệ\ thống\ tóm\ tắt\ tự\ động}$$

➢ Precision: là tỉ lệ các tài liệu được trả về thực sự có liên quan đến tài liệu truy vấn.

Nội dung nghiên cứu

❖ Phương pháp đánh giá (tiếp)

$$F = \frac{Recall \times Precision}{(Recall + Precision)/2}$$

➢ F-score: một độ đo để cân đối tỉ lệ giữa recall và precision.

Nội dung nghiên cứu

❖ Phương pháp đánh giá (tiếp)

$$F = \frac{Recall \times Precision}{(Recall + Precision)/2}$$

➢ F-score: một độ đo để cân đối tỉ lệ giữa recall và precision.

Đánh giá kết quả

- Bộ dữ liệu thực nghiệm bao gồm 5144 bài báo được lấy tự động từ nhiều lĩnh vực trên trang tuoitre.vn và được đưa vào các hệ thống tóm tắt tự động:
  - S1: Sử dụng thuật toán TextRank gốc.
  - S2: Sử dụng thuật toán TextRank kết hợp với độ đo Cosine
  - S3: Sử dụng thuật toán TextRank kết hợp với độ đo Euclidean
  - S4: Tính độ tương tự giữa hai câu bằng tổng số từ đơn giống nhau và kết hợp với TextRank
  - S4: Tính độ tương tự giữa hai câu bằng tổng số cụm từ giống nhau và kết hợp với TextRank

Đánh giá kết quả

❖ Kết quả đánh giá được tính theo phương pháp ROUGE

Hệ thống	Recall	Precision	F-score
S1	0.63	0.26	0.36
S2	0.3	0.22	0.25
S3	0.31	0.21	0.25
S4	0.64	0.26	0.37
S5	0.35	0.16	0.22

Kết Luận

❖ Những điều đã làm được:

- Trình bày hoàn chỉnh mô hình của một hệ thống tóm tắt văn bản tự động.
- Áp dụng được mô hình đồ thị để giải quyết bài toán tóm tắt văn bản Tiếng Việt.
- Đưa ra nhiều phương pháp khác nhau tính toán độ tương đồng giữa hai câu từ đó xây dựng ra các hệ thống tóm tắt khác nhau
- Phương pháp không hạn chế loại dữ liệu cũng như lĩnh vực áp dụng.

Kết Luận

❖ Những mặt còn hạn chế:

- Vấn đề ngữ nghĩa của văn bản tóm tắt.
- Các phương pháp tính độ tương đồng ngữ nghĩa bổ sung chưa đạt kết quả tốt như phương pháp TextRank gốc.
- Chưa áp dụng được các yếu tố về mặt ngữ nghĩa, cú pháp từ vựng để cải thiện chất lượng các văn bản tóm tắt.

Kết Luận

❖ Hướng phát triển:

- Mở rộng không gian tập dữ liệu để cải thiện kết quả của độ đo cosine, Euclidean.
- Ứng dụng các phương pháp để cải thiện việc xác định độ tương đồng giữa hai câu.
- Cải thiện chất lượng của các văn bản tóm tắt.
- Hướng tới phát triển mô hình tóm tắt đa văn bản.


thank
you!