
Text Summarization Project

Cyber Chanh Vo

Tóm tắt văn bản

Sách, báo, Twitter, Facebook, Youtube... là các nguồn thông tin vô tận, nếu tận dụng hiệu quả có thể mang đến lượng tri thức vô cùng lớn.

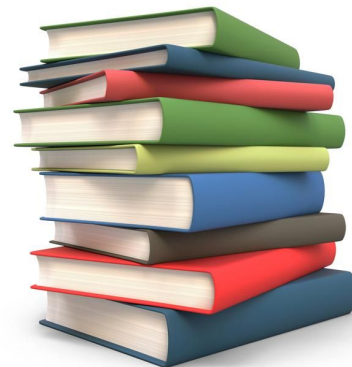
Có thể nhận thấy một cơ chế tóm tắt thông tin nhanh chóng và hiệu quả sẽ hỗ trợ con người chuyển những thông hiện hữu thành thông tin có ích, đẩy nhanh quá trình tiến đến tri thức.



Định nghĩa

Tóm tắt văn bản là quá trình rút trích những thông tin quan trọng nhất từ **một hoặc nhiều nguồn** nhằm tạo ra phiên bản cô đọng, ngắn gọn phục vụ cho **một hoặc nhiều người dùng** cụ thể, hay **một hoặc nhiều nhiệm vụ** cụ thể.

“The process of distilling the most important information from a source (or sources) to produce an abridged version for a particular user (or users) and task (or tasks).” [Mani 2001]





Các ứng dụng hiện tại

- Tóm tắt tin tức
- Hỗ trợ điều trị
- Tóm tắt Result trong search engine
- Thu thập dữ liệu
- Tóm tắt bài báo khoa học
- Tóm tắt nội dung cuộc họp, hội nghị
- Tóm tắt nội dung video, audio
- Trả lời tự động



Phân loại tóm tắt

1. Theo kết quả (output)

→ Rút trích (Extract)

Chứa nội dung được rút trích từ văn bản gốc.

→ Tóm lược (Abstract)

Chứa nội dung không được thể hiện trong văn bản gốc.



2. Theo mục đích (function)

→ Chỉ thị (Indicative)

Cung cấp chức năng tham khảo để chọn tài liệu đọc chi tiết hơn (ứng dụng trong Tóm tắt kết quả tìm kiếm)

VD: Tóm tắt tin tức, tóm tắt đưa ra chi tiết chính của từng sự kiện.

→ Thông tin (Information)

Bao gồm tất cả các thông tin nổi bật có trong văn bản gốc tại nhiều mức độ chi tiết khác nhau.

→ Đánh giá (Evaluation)

Đánh giá vấn đề chính của văn bản nguồn, thể hiện quan điểm của tác giả đối với công việc của họ.



3. Theo nội dung

→ Tóm tắt chung (Generalized)

Đưa ra các nội dung quan trọng bao quát văn bản gốc.

→ Tóm tắt hướng truy vấn (Query-based)

Đưa ra kết quả dựa vào câu truy vấn. Tóm tắt này thường sử dụng trong quá trình tìm kiếm thông tin (Information retrieval)



4. Theo miền dữ liệu

- **Trên 1 miền dữ liệu (Domain)**
Tóm tắt nhắm vào một miền nội dung nào đó, như tin tức khủng bố, tin tức tài chính...
- **Trên 1 thể loại (Genre)**
Nhắm vào một thể loại văn bản nào đó, như báo chí, email, web, bài báo...
- **Tóm tắt độc lập (Independent)**
Tóm tắt cho nhiều thể loại và nhiều miền dữ liệu.



5. Theo mức độ chi tiết

- **Tóm tắt tổng quan (overview)**
Miêu tả tổng quan tất cả các nội dung nổi bật trong văn bản nguồn.
- **Tóm tắt tập trung sự kiện (event)**
Tả một sự kiện cụ thể nào đó trong văn bản nguồn.

6. Theo số lượng

- **Tóm tắt đơn văn bản**
- **Tóm tắt đa văn bản**

7. Theo ngôn ngữ

- **Tóm tắt đơn ngôn ngữ**
- **Tóm tắt đa ngôn ngữ**
- **Tóm tắt xuyên ngôn ngữ (cross-language)**



Đặc điểm của các bản tóm tắt

- **Giảm nội dung thông tin**
Được đo theo Tỷ lệ nén hoặc Chiều dài mong muốn.
- **Nội dung thông tin**
Trung thực và phù hợp với yêu cầu của người dùng
- **Định dạng tốt**
Có cấu trúc và đọc hiểu được.



Các phương pháp đánh giá

→ Đánh giá thủ công

Chuyên gia trực tiếp đánh giá.

→ Đánh giá đồng chọn

Chỉ dùng với tóm tắt dạng trích rút, so sánh văn bản tóm tắt với văn bản gốc theo Precision, Recall và F-score

→ Đánh giá trên nội dung

Phương pháp ROUGE

(Recall-Oriented Understudy for Gisting Evaluation)

Hiện nay được coi như một phương pháp đáng tin cậy để đánh giá độ chính xác của một hệ thống tóm tắt văn bản tự động.

Phương pháp LCS

(Longest Common Subsequence) Độ dài chuỗi con chung dài nhất của 2 văn bản.

Phương pháp BLEU

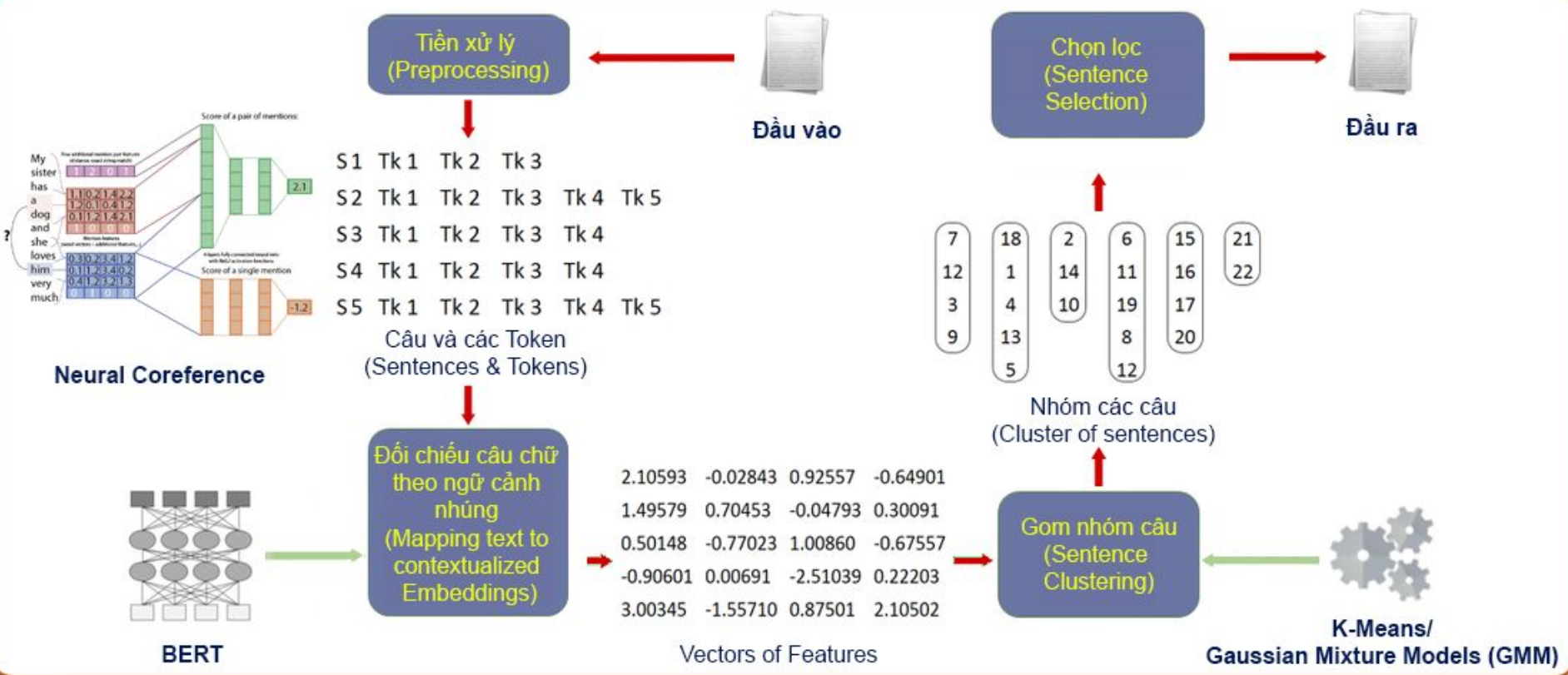
(Bilingual Evaluation Understudy). Đánh giá độ chính xác của hệ thống dịch máy.

Tiếp cận vấn đề

Mô hình (model) để giải quyết bài toán này chủ yếu xoay quanh phương pháp tóm tắt rút trích (Extraction)

Dựa vào BERT model của Transformers để tạo ra các mẫu nhúng (embeddings) phục vụ cho quá trình gom nhóm (clustering), và sử dụng K-Means model để tạo ra một bản tóm tắt.

Cụ thể hơn, model tận dụng một pipeline mã hóa (tokenized) đoạn văn thành các câu rõ ràng, các câu này sẽ được chuyển đến mô hình BERT để suy luận cho các bản nhúng đầu ra. Sau đó gom nhóm các bản nhúng bằng K-Means, chọn ra các câu gần với trọng tâm nhất.



Neural Coreference

Để BERT hoạt động tốt, ta cần quan tâm nhiều đến vấn đề ngữ cảnh, quá trình liên tưởng hay liên kết các mối liên hệ với nhau gọi là **phân giải coreference (tham chiếu lỗi/tham chiếu thành phần chính).**

Ví dụ: Trong các ngữ cảnh khác nhau, từ "nó" có thể là những ai, là thứ gì?

Con người có thể trả lời một cách tự nhiên, nhưng với AI thì điều đó khó hơn nhiều. Một thuật toán phân giải coreference điển hình có các bước như sau:

- *Trích xuất các đề cập (mentions) và những từ có khả năng liên hệ đến đề cập, những từ này gọi là các tiền đề (antecedent)*
- *Với mỗi mention và bộ từ liên quan, ta tính ra được một bộ tính năng (features). Dựa trên bộ feature này, tìm ra antecedent có khả năng xảy ra nhất cho mỗi mention.*
- *Cuối cùng là xếp hạng theo bộ.*

Hiển nhiên, các bộ feature từ đặc trưng ngôn ngữ có thể rất lớn, một số ngôn ngữ phức tạp có thể lên đến hơn 120 feature, việc setup thủ công các bộ feature này gây hao phí nhân lực và thời gian.

→ **Các kỹ thuật NLP như Word vector và mạng Neural** có thể giải quyết được vấn đề này. Chúng có thể tự động học hỏi nhiều feature thủ công, giảm feature theo xếp hạng trong khi vẫn giữ được độ chính xác lý tưởng, thậm chí tốt hơn.

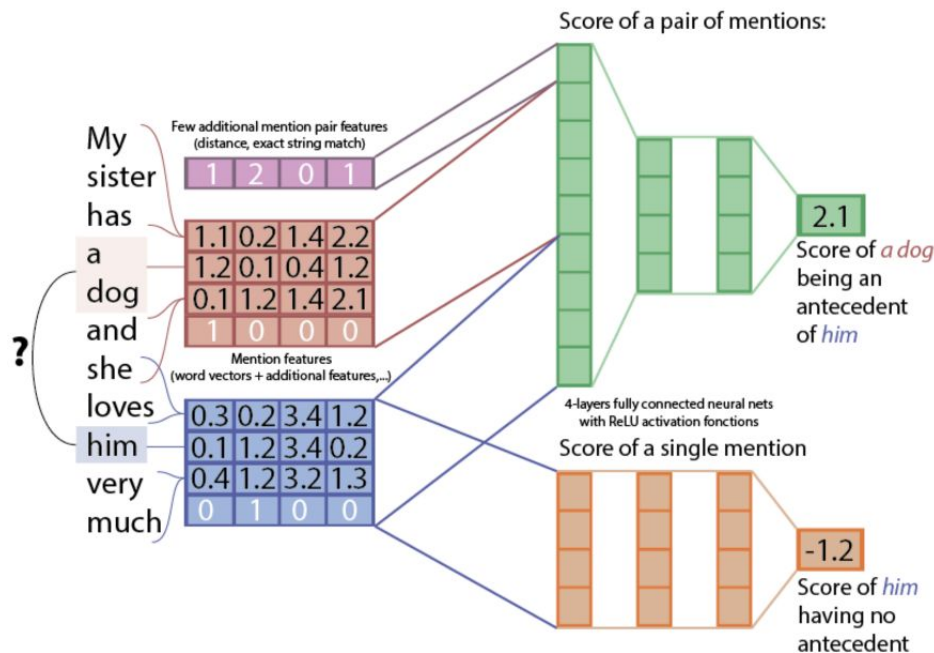
Pipeline hiện tại của Neural Coreference dựa trên thư viện **Numpy** kết hợp với **khả năng phân tích cú pháp tốc độ cao** của **thư viện ngôn ngữ spaCy**.

Với khả năng phát hiện mention, trích xuất feature và tính toán neural, nó có thể dùng những định nghĩa, thông tin được cung cấp để **tính toán nhanh phần nhúng (embedding) cho các từ không xác định**.

Neural Coreference

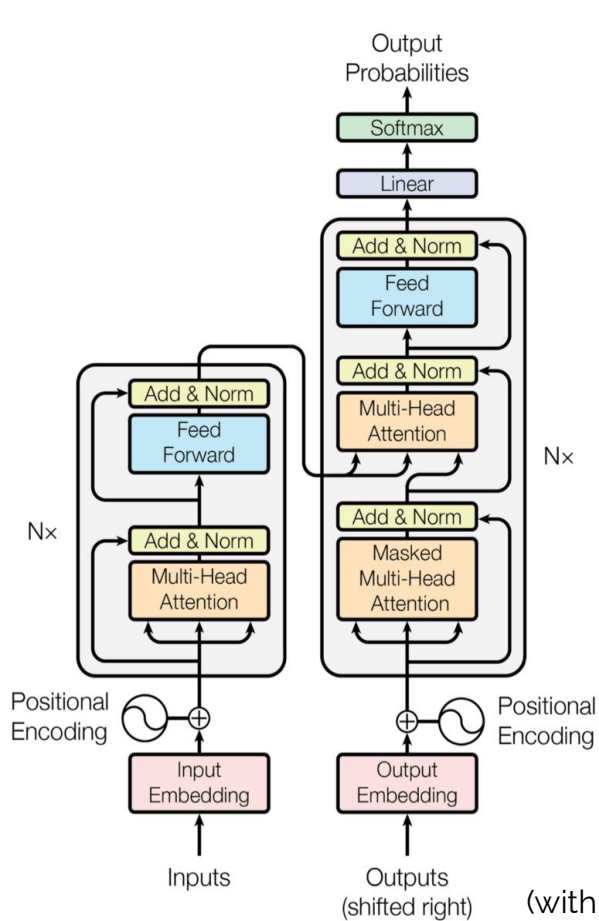
Model tổng quát:

- Nhúng một số từ bên trong và xung quanh mỗi đề cập (mention), tính trung bình chúng nếu cần.
- Thêm một số feature đơn giản (độ dài của đề cập, thông tin người nói, vị trí của đề cập...) để lấy feature đại diện cho mỗi đề cập và môi trường xung quanh nó.
- Đưa các biểu diễn này vào 2 mạng neural.
- **Mạng đầu tiên cho điểm cho mỗi cặp đề cập và tiền đề (antecedent) có thể có trong khi mạng thứ hai cho điểm cho một đề cập không có tiền đề** (đôi khi đề cập là chủ thể đầu câu trong văn bản).

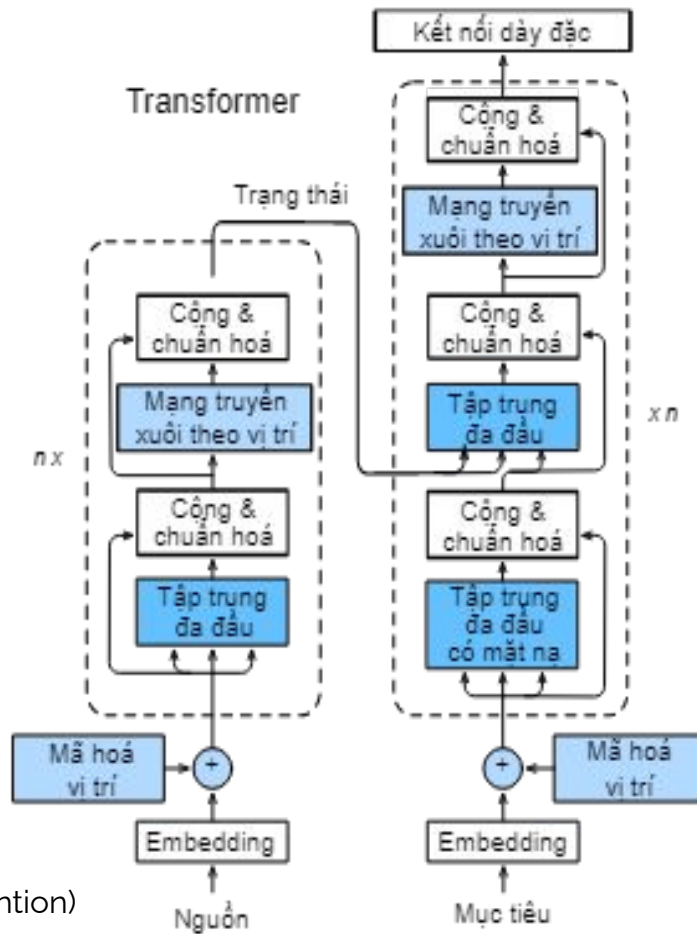


Sau đó, chỉ cần so sánh tất cả các điểm này với nhau, lấy điểm cao nhất để xác định **một đề cập có tiền đề hay không và nó nên là tiền đề nào**.

BERT from Transformers



Transformers
(with Multi-head Attention)



BERT from Transformers

Bidirectional Encoder Representation from Transformer nghĩa là **model biểu diễn từ** theo 2 chiều ứng dụng kỹ thuật **Transformer**, được thiết kế để tiền huấn luyện các phép nhúng từ (**pre-train word embedding**).

Tài liệu về BERT khá nhiều nên ta chỉ nói một cách đơn giản về cách thức hoạt động.

Cơ chế Attention của Transformer sẽ truyền toàn bộ câu vào mô hình cùng lúc, không quan tâm đến chiều của câu. Do đó model này được xem như là huấn luyện 2 chiều (bidirectional) dù có thể nói đó là huấn luyện không chiều (non-directional).

Đặc điểm này cho phép nó học được bối cảnh của từ dựa trên toàn bộ các từ xung quanh nó, bao gồm cả từ bên trái và bên phải. Các khả năng đặc trưng của BERT:

1. **Có thể Fine-tuning**, đây là đặc điểm các model embedding trước đây chưa làm được.
2. **Masked ML (MLM)**
3. **Next Sentence Prediction (NSP)**

BERT from Transformers

Có khá nhiều phiên bản khác nhau của BERT, dựa trên việc thay đổi kiến trúc của Transformer tập trung ở 3 tham số:

- *L: số lượng block sub-layers trong Transformer*
- *H: kích thước của embedding vector (hidden size)*
- *A: số lượng head trong multi-head layer, mỗi một head sẽ thực hiện một self-attention.*

BERT base (L = 12, H = 768, A = 12): 110 triệu tham số

BERT large (L = 24, H = 1024, A = 16): 340 triệu tham số

BERT multilingual (L = 12, H = 768, A = 12, 102 ngôn ngữ): 168 triệu tham số

Các model khác được sử dụng trong bài:

GPT(12-768-12: 110M), **GPT2** (12-768-12: 117M), **BART large** (24-1024-16: 406M), **Transformer XL**(18-1024-16: 257M), **XLNet large**(24-1024-16: 340M), **DistilBERT**(6-768-12: 66M), **CTRL**(48-1280-16: 1.6B), **ALBERT large v2**(24 repeating layers, 128 embedding, H = 1024, A = 16: 17M), **phoBERT**(350M), ...

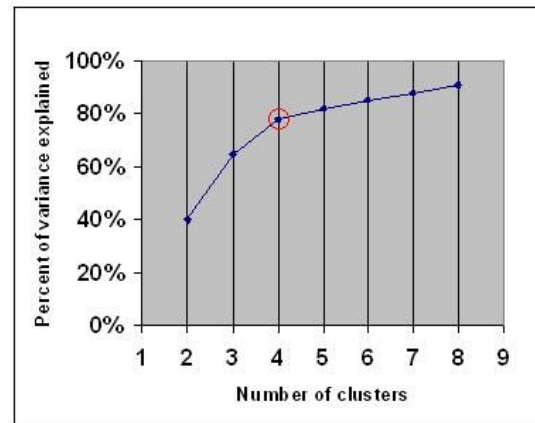
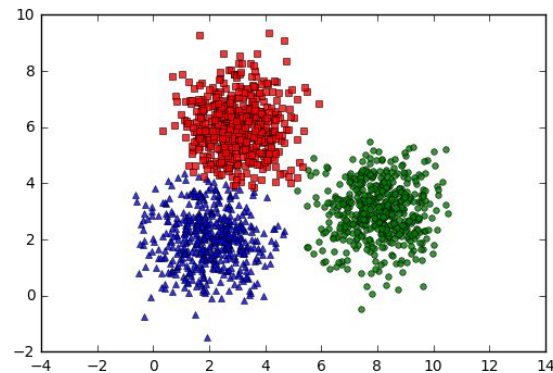
K-Means

Elbow method

Về cơ bản, thuật toán **K-Means** đơn giản và khá hiệu quả trong việc gom nhóm các câu phức vụ đầu ra.

Yêu cầu chính của K-Means là **số lượng K cluster cần phải xác định**, lúc này ta áp dụng **Elbow method** để tìm K.

Dựa vào biểu đồ có dạng “cùi chỏ”, ta thấy **Elbow method** có thể hiểu đơn giản là tìm số cụm tối ưu, sao cho việc thêm một cụm khác không mang lại model tốt hơn đáng kể.



Sample code

Các thư viện cần cài đặt

werkzeug

bert-extractive-summarizer

jupyter-client==6.1.5

wrapt==1.12.1

traitlets==5.0

typed-ast

pytest-filter-subpackage

pytest-cov

tensorflow --user

pyvi

numpy

torch

spacy==2.1.3

transformers==3.3.0

sentencepiece

Cython

tqdm==4.32.2

neuralcoref --no-binary neuralcoref

argparse

scikit-learn

pytest

rouge-metric



Phần mềm sử dụng:

Jupyter Notebook, Python 3.7

**Python 3.9 không tương thích với một số thư viện*

Cấu hình máy local:

ASUS GL552VX, I5 6300HQ

GTX950M, 32GB RAM

Classes & Functions

Với body là văn bản đầu vào, ta có cách gọi đơn giản như sau:

```
model = Summarizer() #model = Summarizer("bert-large-uncased")
```

```
result = model(body)
```

class Summarizer: gọi model chạy nhiệm vụ chính của bài toán

- **model:** tên model, mặc định 'bert-base-multilingual-uncased'
- *custom_model:* mặc định là None, khi gán sẽ override model
- *custom_tokenizer:* mặc định là None, khi gán sẽ override model tokenizer
- *hidden:* chọn layer embeddings
- *reduce_option:* thu gọn kết quả, mặc định là trung bình cộng 'mean'
- **sentence_handler:** xử lý ngữ cảnh, mặc định là **SentenceHandler()**
- *random_state:* trạng thái ngẫu nhiên để tạo bản tóm tắt
- *hidden_concat:* ghép nối nhiều hidden layer, mặc định là không

Classes & Functions

Ví dụ:

```
model = Summarizer()
```

```
result = model(body, ratio = 0.5) #result = model(body, num_sentences = 5)
```

class ModelProcessor: thực thi nhiệm vụ chính của bài toán như preprocess các câu, nhúng (embeddings) (gọi class **BertParent**), gom nhóm (clustering) (gọi class **ClusterFeatures**), truy xuất các đoạn tóm tắt đã nhúng (retrieve summarized embedding)

- **body:** Đầu vào
- **ratio:** Độ nén, tỉ lệ giữa đầu ra/đầu vào, mặc định là 0.2
- **min_length:** Độ dài tối thiểu cho các câu dùng để tóm tắt, mặc định 40
- **max_length:** Độ dài tối đa, mặc định 600
- **use_first:** Sử dụng câu đầu tiên, mặc định là có
- **algorithm:** Thuật toán gom nhóm (mặc định K-Means hoặc Gaussian Mixture)
- **num_sentences:** Số câu của đầu ra, mặc định là None, nếu nhập sẽ override ratio.
- **return_as_list:** Kết quả trả ra dạng list, mặc định là không

Classes & Functions

Ví dụ:

```
model = Summarizer(sentence_handler=CoreferenceHandler(greedyneess=0.4))
```

```
result = model(body, ratio = 0.5)
```

class SentenceHandler: xử lý ngữ cảnh mặc định bằng thư viện ngôn ngữ.

class CoreferenceHandler: xử lý ngữ cảnh bằng thư viện ngôn ngữ và Neural Coreference.

Model Neural Coreference chuẩn có rất nhiều tham số, ở bài toán này ta chỉ sử dụng greedyneess (tham lam) có giá trị từ 0-1, mặc định là 0.45, số càng lớn thì càng nhiều tham chiếu đến lỗi.

Cả 2 class xử lý ngữ cảnh trên đều mặc định dùng thư viện tiếng Việt của spaCy (from spacy.lang.vi import Vietnamese)

class ClusterFeatures: thực hiện gom nhóm (clustering) và các function hỗ trợ (Elbow method, tìm tâm,...)

class BertParent: quản lý model (BERT/ XLNet, ...). Gồm 2 nhiệm vụ chính là mã hóa đầu vào (tokenize input) và tạo ma trận nhúng (extract embeddings, create matrix)

Classes & Functions

def rouge_dist(hypotheses, references): gọi phép đo Rouge để đánh giá kết quả.

Bộ thư viện tương ứng đòi hỏi input khá chi tiết.

hypotheses là **list of str** - [STR1, STR2, ...], các văn bản tóm tắt đầu ra của model.

references là **list of list of str** - [[str1a, str1b], [str2a, str2b], ...], các văn bản tóm tắt mẫu.

2 list phải có độ dài bằng nhau.

*Khi check từng phần tử sẽ là quan hệ **STR1 vs [str1a, str1b]**. Điều này có thể phát sinh yêu cầu tinh chỉnh format lúc nạp các văn bản tóm tắt mẫu vào model.*

*Nói đơn giản là **mỗi văn bản tóm tắt của model** sẽ được so sánh với **1 list văn bản tóm tắt mẫu tương ứng** và trả ra giá trị trung bình cộng.*

Có thể gọi nhiều phép đo Rouge-N, Rouge-L, ... cùng lúc.

Đầu ra là một dictionary với Precision, Recall, Score ứng với từng loại Rouge.

Các con số cuối cùng là trung bình cộng các kết quả.

Sample

Ví dụ body (1710 từ, khoảng 73 câu), nguồn Wikipedia, đã được loại bỏ hình ảnh và link

```
In [104]: body = '''
Shiba Inu (柴犬 (sài khuyển)) là loại chó nhỏ nhất trong sáu giống chó nguyên thủy và riêng biệt đến từ Nhật Bản. Chúng là một
Shiba là một trong sáu giống chó điển hình của Nhật Bản, cũng như Hokkaido, Kishu, Shikoku, Kai và Akita. Trong những giống chó
Inu hoặc ken (犬 - Hán Việt: khuyển) trong tiếng Nhật có nghĩa là con chó, nhưng nguồn gốc của từ "Shiba" vẫn chưa rõ. Từ Shiba
Khung hình của Shiba nhỏ gọn với cơ bắp phát triển tốt. Con đực có chiều cao từ 35 đến 43 cm (14 đến 17 in). Đối với con cái là
Lớp lông: Có hai lớp lông với lớp ngoài cứng và thẳng cùng một lớp trong mềm mại và dày. Lông mao ngắn và thậm chí trên mặt, ta
Urajiro (màu kem trắng) có ở các bộ phận sau trên tất cả các vùng lông: ở hai bên mõm, trên má, bên trong tai, trên hàm dưới và
Shiba có xu hướng thể hiện tính tự lập và đôi khi còn hung hăng. Shiba Inu tốt nhất nên được nuôi trong một gia đình mà không c
Một tinh thần mạnh mẽ, một bản chất tốt đẹp và sự thẳng thắn không lẫn lộn mang lại phẩm giá và vẻ đẹp tự nhiên. Shiba có tính
Shiba là một giống chó tương đối khó tính và cảm thấy rất cần thiết khi giữ chính nó thật sạch. Nó thường liếm bàn chân giống n
Một đặc điểm giúp phân biệt giống chó này là "Shiba scream". Khi đủ kích động hay không vui, nó sẽ phát ra một tiếng thét lớn v
Thí nghiệm phân tích DNA gần đây đã khẳng định rằng loài chó mõm nhọn châu Á này là một trong những giống chó lâu đời nhất, đã
Ban đầu, Shiba Inu được nuôi để săn và bắt các con vật nhỏ, chẳng hạn như các loài chim và thỏ. Dù đã cố nhiều nỗ lực để bảo tồ
Năm 1954, một gia đình phục vụ vũ trang mang con Shiba Inu đầu tiên đến Hoa Kỳ. Vào năm 1979, lứa đầu tiên được ghi nhận sinh r
Một con Shina Inu đang chơi đùa ở bãi cỏ.
Tình trạng sức khỏe được biết ảnh hưởng đến giống chó này là dị ứng, thanh quang nhãn, cườm thủy tinh thể mắt, loạn sản xương h
Kiểm tra chung định kỳ được khuyến cáo nên được thực hiện trong suốt cuộc đời của con chó nhưng vấn đề thường được phát hiện số
Nhu đối với bất kỳ những con chó khác, Shiba nên được đi hoặc nếu không thì nên vận động hàng ngày.
Tuổi thọ trung bình của Shiba Inu là từ 12 đến 16 năm. Tập thể dục, đặc biệt là đi bộ mỗi ngày, sẽ giúp cho giống chó này sống
Giống chó này rất sạch sẽ, vì vậy nhu cầu chải chuốt nên được thực hiện tối thiểu. Một lớp lông Shiba Inu thô, ngắn có chiều dài
'''
```

executed in 11ms, finished 12:43:08 2021-07-22

Sample

Model mặc định (BERT multilingual), số câu tóm tắt là 10.

```
#default model with number of senteces = 10
model = Summarizer()
result = model(body, num_sentences = 10)
print(result)
```

executed in 11.6s, finished 12:48:59 2021-07-22

Shiba Inu (柴犬 (sài khuyển)) là loại chó nhỏ nhất trong sáu giống chó nguyên thủy và riêng biệt đến từ Nhật Bản. Điều này khiến cho một số người tin rằng Shiba được đặt tên như thế là vì loài chó này được sử dụng để săn mồi trong các bụi cây, hoặc có thể là do màu sắc phổ biến nhất của Shiba Inu là màu đỏ tương tự như của các cây bụi. Đối với con cái là 33 đến 41 cm (13 đến 16 in). Shiba có thể có màu đỏ, đen và nâu, hoặc màu vừng (màu đỏ với những sợi ngà sang đen), với một lớp lông lót màu kem, màu da bò, hoặc màu xám. Urajiro (màu kem trắng) có ở các bộ phận sau trên tất cả các vùng lông: ở hai bên mõm, trên má, bên trong tai, trên hàm dưới và ở chỗ cổ họng, bên trong chân, trên bụng, xung quanh các lỗ thông hơi và phía vùng bụng của đuôi. Shiba có tính chất tự lập và có thể dè dặt đối với người lạ nhưng lại trung thành và tình cảm với những người có được sự tôn trọng của nó. Nó có thể hung dữ với những con chó khác. Khi nghiên cứu về chó Nhật được chính thức hóa trong đầu và giữa thế kỷ 20, ba chủ ng này đã được kết hợp thành một giống tổng thể, Shiba Inu. Nhìn chung, dù gì đi nữa, chúng có tính di truyền cao và khá nhiều Shiba được chẩn đoán khuyết tật do di truyền so với các giống chó khác. Tuy nhiên, rụng lông có thể là một mối phiền toái.

Model BERT large, số câu tóm tắt là 10.

```
#BERT Large Uncased model with number of senteces = 10
model = Summarizer(model = 'bert-large-uncased')
result = model(body, num_sentences = 10)
print(result)
```

executed in 32.0s, finished 12:47:07 2021-07-22

Shiba Inu (柴犬 (sài khuyển)) là loại chó nhỏ nhất trong sáu giống chó nguyên thủy và riêng biệt đến từ Nhật Bản. Tuy nhiên, trong một phương ngữ Nagano cổ, từ Shiba cũng có ý nghĩa là nhỏ, do đó cái tên có thể nói đến tầm vóc nhỏ bé của con chó. Khung hình của Shiba nhỏ gọn với cơ bắp phát triển tốt. Một tinh thần mạnh mẽ, một bản chất tốt đẹp và sự thẳng thắn không lẫn lộn mang lại phẩm giá và vẻ đẹp tự nhiên. Nó có thể hung dữ với những con chó khác. Nó thường liếm bàn chân giống như mèo, thường đi c huyển theo cách riêng của mình để giữ bộ lông sạch sẽ, nhưng lại cực kỳ thích bơi lội và chơi đùa trong các vùng nước. Vì bản c hất khó tính và đầy kiêu hãnh vốn có, Shiba con rất dễ dạy dỗ và trong nhiều trường hợp sẽ tự dạy dỗ chính mình. Giống bây giờ chủ yếu được nuôi như thú cưng ở Nhật Bản và các nước khác. Tập thể dục, đặc biệt là đi bộ mỗi ngày, sẽ giúp cho giống chó này sống lâu và khỏe mạnh. Giống chó này rất sạch sẽ, vì vậy nhu cầu chải chuốt nên được thực hiện tối thiểu. Nó cũng có một lớp lông dày có thể bảo vệ chúng khỏi nhiệt độ đông đá.

Sample

Model BERT large, độ nén 0.3

```
#BERT large Uncased model with ratio = 0.3
model = Summarizer(model = 'bert-large-uncased')
result = model(body, ratio = 0.3)
print(result)

executed in 33.7s, finished 12:47:41 2021-07-22
```

Shiba Inu (柴犬 (sài khuyển)) là loại chó nhỏ nhất trong sáu giống chó nguyên thủy và riêng biệt đến từ Nhật Bản. Inu hoặc ken (犬 - Hán Việt: khuyển) trong tiếng Nhật có nghĩa là con chó, nhưng nguồn gốc của từ "Shiba" vẫn chưa rõ. Khung hình của Shiba Inu gọn với cơ bắp phát triển tốt. Đối với con cái là 33 đến 41 cm (13 đến 16 in). Giống chó cũng tương tác khá tốt với mèo. Nó có thể hung dữ với những con chó khác. Nó thường liếm bàn chân giống như mèo, thường đi chuyển theo cách riêng của mình để giữ bộ lông sạch sẽ, nhưng lại cực kỳ thích bơi lội và chơi đùa trong các vùng nước. Vì bản chất khó tính và đầy kiêu hãnh vốn có, Shiba Inu con rất dễ dạy dỗ và trong nhiều trường hợp sẽ tự dạy dỗ chính mình. Một đặc điểm giúp phân biệt giống chó này là "Shiba scream". Những dòng máu đó là Shinshu Shiba từ Nagano, Mino Shiba từ Gifu, và San'in Shiba từ Tottori và Shimane. San'in Shiba thì lớn hơn so với hầu hết các giống Shiba hiện nay, và thường có màu đen, không có dấu sẫm và trắng thường được tìm thấy trên Shiba đen - sẫm hiện nay. Vào năm 1979, lứa đầu tiên được ghi nhận sinh ra tại Hoa Kỳ. Giống bây giờ chủ yếu được nuôi như thú cưng ở Nhật Bản và các nước khác. Kiểm tra chung định kỳ được khuyến cáo nên được thực hiện trong suốt cuộc đời của con chó như ng vấn đề thường được phát hiện sớm trong cuộc đời của nó. Năm hai tuổi, Shiba Inu có thể được coi là hoàn toàn tự do khỏi các vấn đề chung nếu không được phát hiện bởi thời điểm này, vì ở độ tuổi này bộ xương đã được phát triển đầy đủ. Như đối với bất kỳ những con chó khác, Shiba nên được đi hoặc nếu không thì nên vận động hàng ngày. Tập thể dục, đặc biệt là đi bộ mỗi ngày, sẽ giúp cho giống chó này sống lâu và khỏe mạnh. Giống chó này rất sạch sẽ, vì vậy nhu cầu chải chuốt nên được thực hiện tối thiểu. Nó cũng có một lớp lông dày có thể bảo vệ chúng khỏi nhiệt độ đông đá. Tuy nhiên, rụng lông có thể là một mối phiền toái. Rụng lông nặng nhất có sự thay đổi theo mùa và đặc biệt là trong mùa hè, nhưng việc chải lông hàng ngày có thể làm giảm vấn đề này.

Model BERT large, độ nén 0.3,
sentence handler là Neural
coreference

```
#default model with CoreferenceHandler
model = Summarizer(model = 'bert-large-uncased', sentence_handler=CoreferenceHandler())
result = model(body, ratio = 0.3)
print(result)

executed in 36.6s, finished 12:48:18 2021-07-22
```

Shiba Inu (柴犬 (sài khuyển)) là loại chó nhỏ nhất trong sáu giống chó nguyên thủy và riêng biệt đến từ Nhật Bản. Chúng là một giống chó nhỏ, nhanh nhẹn và thích hợp với địa hình miền núi, Shiba Inu ban đầu được nuôi để săn bắt. Inu hoặc ken (犬 - Hán Việt: khuyển) trong tiếng Nhật có nghĩa là con chó, nhưng nguồn gốc của từ "Shiba" vẫn chưa rõ. Khung hình của Shiba Inu gọn với cơ bắp phát triển tốt. Con đực có chiều cao từ 35 đến 43 cm (14 đến 17 in). Trọng lượng trung bình ở kích thước trưởng thành là khoảng 10 kg (22 lb) đối với con đực và 8 kg (18 lb) đối với con cái. Giống chó cũng tương tác khá tốt với mèo. Nó có thể hung dữ với những con chó khác. Nó thường liếm bàn chân giống như mèo, thường đi chuyển theo cách riêng của mình để giữ bộ lông sạch sẽ, nhưng lại cực kỳ thích bơi lội và chơi đùa trong các vùng nước. Một đặc điểm giúp phân biệt giống chó này là "Shiba scream". Khi đủ kích động hay không vui, nó sẽ phát ra một tiếng thét lớn và cao. Năm 1954, một gia đình phục vụ vũ trang mang con Shiba Inu đầu tiên đến Hoa Kỳ. Vào năm 1979, lứa đầu tiên được ghi nhận sinh ra tại Hoa Kỳ. Giống bây giờ chủ yếu được nuôi như thú cưng ở Nhật Bản và các nước khác. Kiểm tra chung định kỳ được khuyến cáo nên được thực hiện trong suốt cuộc đời của con chó nhưng vấn đề thường được phát hiện sớm trong cuộc đời của nó. Năm hai tuổi, Shiba Inu có thể được coi là hoàn toàn tự do khỏi các vấn đề chung nếu không được phát hiện bởi thời điểm này, vì ở độ tuổi này bộ xương đã được phát triển đầy đủ. Như đối với bất kỳ những con chó khác, Shiba nên được đi hoặc nếu không thì nên vận động hàng ngày. Tập thể dục, đặc biệt là đi bộ mỗi ngày, sẽ giúp cho giống chó này sống lâu và khỏe mạnh. Giống chó này rất sạch sẽ, vì vậy nhu cầu chải chuốt nên được thực hiện tối thiểu. Tuy nhiên, rụng lông có thể là một mối phiền toái. Rụng lông nặng nhất có sự thay đổi theo mùa và đặc biệt là trong mùa hè, nhưng việc chải lông hàng ngày có thể làm giảm vấn đề này.

Sample

Độ nén 0.3, ma trận nhúng (embeddings)

```
result = model.run_embeddings(body, ratio=0.3) # Will return (num_sentences +1, N) embedding numpy matrix.  
result
```

executed in 10.7s, finished 12:52:13 2021-07-22

```
array([[ 0.06264511, -0.04616188,  0.30910435, ..., -0.11081592,  
        -0.06871919,  0.11080622],  
       [ 0.0010872 , -0.0021876 ,  0.04671476, ...,  0.42539126,  
        -0.261949 ,  0.09410372],  
       [-0.03221155, -0.21116859,  0.19516714, ..., -0.12356774,  
        -0.13606201,  0.36228496],  
       ...,  
       [-0.4348117 ,  0.3021262 ,  0.21845956, ..., -0.6307475 ,  
        -0.06835769, -0.02482057],  
       [-0.21573763,  0.32624435, -0.06969851, ..., -0.20903395,  
        -0.34038976, -0.0154977 ],  
       [-0.18740307, -0.29025468, -0.01592006, ..., -0.175454 ,  
        0.24482788,  0.2545929 ]], dtype=float32)
```

```
res = model.calculate_elbow(body, k_max=10)  
res
```

executed in 5.61s, finished 12:52:35 2021-07-22

```
[2692.918701171875,  
 2500.3896484375,  
 2321.54345703125,  
 2202.852783203125,  
 2082.740966796875,  
 2033.31689453125,  
 1945.07568359375,  
 1863.1490478515625,  
 1775.525390625]
```

```
res = model.calculate_optimal_k(body, k_max=10)  
res
```

executed in 5.55s, finished 12:52:40 2021-07-22

Giới hạn K tối đa là 10.

Elbow method và

giá trị K tối ưu tương ứng.

Đánh giá kết quả trên các model



BERT, GPT2, BART,
GPT, CTRL,
Transformer XL,
XLNet, DistilBERT,
ALBERT, SciBERT,
phoBERT

Dữ liệu thực nghiệm

Tập dữ liệu trong đề tài “Nghiên cứu một số phương pháp tóm tắt văn bản tự động trên máy tính áp dụng cho Tiếng Việt” của PGS.TS. Lê Thanh Hương, ĐH Bách Khoa Hà Nội.

Tập dữ liệu gồm 200 văn bản thô và 200 tóm tắt mẫu tương ứng, gồm 6 chủ đề.

Link download dataset:
<https://users.soict.hust.edu.vn/huonglt/donvanban.rar>

Chủ đề	Số văn bản
Khoa học công nghệ	25
Chính trị	31
Khoa học - giáo dục	22
Kinh tế	53
Văn hóa	34
Xã hội	35

Đánh giá kết quả

Độ dài văn bản tóm tắt của model được giới hạn theo độ nén (Ratio) gần tương đương với độ dài văn bản mẫu do người tóm tắt.

Dữ liệu được đánh giá theo **phương pháp** (hay độ đo) **ROUGE** với các tiêu chí:

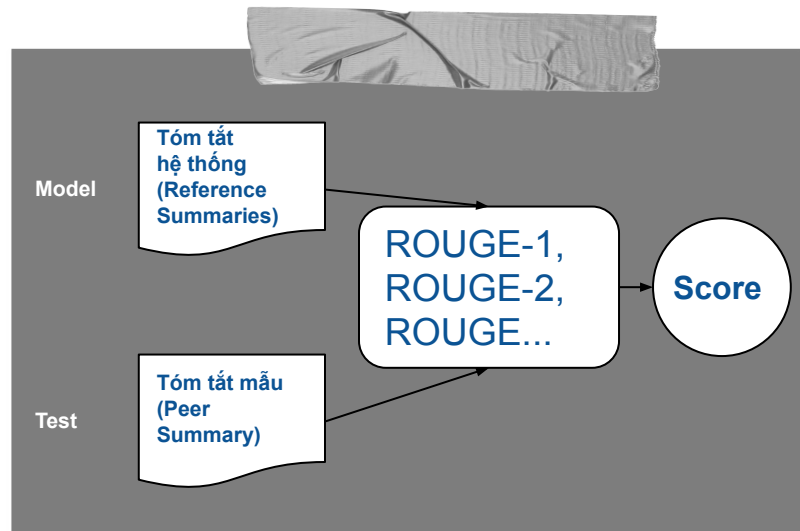
- Đánh giá toàn bộ văn bản trong mỗi bộ dữ liệu.
- Sử dụng ROUGE-1, 2, ROUGE-L và ROUGE-SU4
- Bao gồm cả từ dừng (stop words) trong đánh giá.
- Kết quả đánh giá là kết quả trung bình của bộ dữ liệu.

ROUGE

(Recall-Oriented Understudy for Gisting Evaluation)

Nói một cách đơn giản nó là số liệu để đo lường chất lượng của bản tóm tắt (hoặc bản dịch) của tài liệu tham khảo.

Với những biến thể được sử dụng rộng rãi như R-N, R-L, R-W, R-S, R-SU, ...



Đánh giá kết quả

- *ROUGE-N*

Tính toán các cụm N từ (liền nhau) giống nhau giữa các mẫu thử.

- *ROUGE-L*

Tính toán chuỗi con chung dài nhất bằng **LCS**. LCS ưu tiên thứ tự các từ để phản ánh cấp độ trong câu hơn là sự liên tiếp. Nó tìm ra chuỗi con gồm N từ chung dài nhất, nên không cần N được xác định trước.

- *ROUGE-SU*

Là kết hợp giữa **Rouge-S** (Skip bigram - kết hợp các từ cách xa nhau) và **Rouge-1** (Unigram). Trong trường hợp Rouge-S = 0 thì đánh giá vẫn đạt được giá trị từ Rouge-1. Rouge-SU4 nghĩa là giới hạn bước nhảy tối đa 4 từ.

**Precision cao: Độ chính xác của các kết quả tìm được cao. Tránh False Positive (vd bài toán check Spam)*

**Recall cao: Tỷ lệ bỏ sót các sample positive thực thấp. Tránh False Negative (vd bài toán chẩn đoán Cancer)*

**Bài toán Extraction sẽ quan tâm Recall và Score hơn để tránh bỏ sót các từ khóa quan trọng của data gốc.*

Đánh giá

Kết quả chung

Ta thấy **BERT, phoBERT, XLNet** cho kết quả lý tưởng nhất trên toàn bộ 6 chủ đề tiếng Việt.

Nghĩa là các model pretrain trên **tiếng Việt** (phoBERT) và **đa ngôn ngữ** (BERT multilingual và XLNet) **có lợi thế hơn**, vì BERT nguyên bản được pretrain trên tiếng Anh.


	Rouge1	Rouge2	RougeL	RougeSU4
	f (Score)			
BERT multilingual uncased	0.431400751	0.312293094	0.338213946	0.294229161
BERT large uncased	0.393607677	0.280329732	0.303810227	0.26474728
GPT2	0.40805545	0.298513689	0.321450253	0.280770858
BART large (fb)	0.408656975	0.292236557	0.317803978	0.276233828
GPT (openAI)	0.41055211	0.295769624	0.319011678	0.278949856
CTRL	0.423563082	0.302257521	0.321007755	0.286193332
TransformerXL	0.387149756	0.277917278	0.30213882	0.261223541
XLNet large	0.429643413	0.312409603	0.334935451	0.295496731
DistilBERT base uncased	0.415374496	0.297417118	0.324360531	0.281742816
ALBERT large v2	0.395786381	0.280627994	0.305733132	0.26479345
SciBERT uncased (allenAI)	0.407478271	0.289453472	0.312740254	0.271788329
phoBERT large (vinAI)	0.430087505	0.308554198	0.332093897	0.289732156


Các model khác dù ưu việt hay cải tiến hơn BERT nhưng lại phục vụ chuyên cho các nhu cầu khác (dịch thuật, phân loại, đoán cảm xúc ...) nên kết quả chung không cao.

Với các bảng kết quả tóm tắt theo từng chủ đề cụ thể, thì một số model cải tiến thể hiện ưu thế hơn, ví dụ như **Văn hóa** (GPT), **Xã hội** (BART, CTRL)...

Đánh giá

Kết quả tóm tắt theo từng chủ đề


 Giá trị cao nhất trong chủ đề này


 Giá trị cao nhất trong cả 6 chủ đề

		Rouge1	Rouge2	RougeL	RougeSU4	Rouge1	Rouge2	RougeL	RougeSU4	Rouge1	Rouge2	RougeL	RougeSU4
		p (Precision)				r (Recall)				f (Score)			
Công nghệ	BERT multilingual uncased	0.5899493	0.4688147	0.4714248	0.4400704	0.5316497	0.4192021	0.4220206	0.3926858	0.5592843	0.4426225	0.4453567	0.41503
	BERT large uncased	0.5774234	0.4420538	0.440485	0.4133166	0.4430222	0.3377614	0.3364074	0.3152938	0.501372	0.3829336	0.3814748	0.3577115
	GPT2	0.5928371	0.456343	0.4592389	0.4237097	0.4774641	0.3660816	0.3673619	0.3394596	0.5289323	0.4062592	0.4081943	0.3769342
	BART large (fb)	0.5591487	0.4180205	0.4308391	0.3884065	0.4725079	0.3580276	0.3659995	0.3330335	0.5121902	0.3857052	0.3957813	0.358595
	GPT (openAI)	0.5722911	0.4335628	0.4426835	0.4045676	0.4667793	0.3556946	0.3617078	0.3316068	0.5141781	0.3907875	0.3981199	0.3644717
	CTRL	0.5721954	0.4393822	0.4305467	0.4131478	0.5168073	0.3889703	0.3822331	0.3636264	0.5430928	0.4126423	0.4049539	0.3868085
	TransformerXL	0.5994718	0.4542979	0.4460871	0.4236467	0.4119459	0.3042408	0.3028317	0.2827942	0.4883244	0.3644269	0.3607583	0.3391787
	XLNet large	0.585887	0.4633784	0.4665257	0.4352208	0.5325941	0.4180542	0.4195478	0.3921261	0.5579709	0.439551	0.4417914	0.4125511
	DistilBERT base uncased	0.5620642	0.4239515	0.431258	0.3957797	0.4951833	0.3791243	0.3834639	0.3543351	0.5265084	0.4002868	0.4059591	0.3739124
	ALBERT large v2	0.5764954	0.4448512	0.4461794	0.4181324	0.4610667	0.3563491	0.3572757	0.3345472	0.5123603	0.3957121	0.3968089	0.3716987
	SciBERT uncased (allenAI)	0.5568422	0.4181616	0.4246239	0.3892335	0.4866412	0.3668892	0.3706127	0.3420961	0.5193803	0.3908511	0.3957841	0.3641457
	phoBERT large (vinAI)	0.5562047	0.4323349	0.43073	0.4013904	0.5263373	0.409172	0.4065882	0.3799191	0.540859	0.4204347	0.4183111	0.3903597

Đánh giá

Kết quả tóm tắt theo từng chủ đề

 Giá trị cao nhất trong chủ đề này

 Giá trị cao nhất trong cả 6 chủ đề

		Rouge1	Rouge2	RougeL	RougeSU4	Rouge1	Rouge2	RougeL	RougeSU4	Rouge1	Rouge2	RougeL	RougeSU4
		p (Precision)				r (Recall)				f (Score)			
Chính trị	BERT multilingual uncased	0.6068981	0.4884257	0.5113535	0.4643207	0.4609909	0.3722693	0.3872331	0.3526735	0.5239767	0.4225095	0.440721	0.4008685
	BERT large uncased	0.5847289	0.4664075	0.4935963	0.4472454	0.4235222	0.3423727	0.3571761	0.3279261	0.4912381	0.3948791	0.4144488	0.3784026
	GPT2	0.5918838	0.4703657	0.4950387	0.446783	0.4285297	0.3473095	0.3613326	0.3296259	0.4971313	0.3995779	0.4177478	0.3793652
	BART large (fb)	0.5716417	0.4497268	0.4749025	0.4306904	0.4411813	0.354861	0.3713114	0.3397614	0.4980093	0.3967013	0.4167663	0.3798601
	GPT (openAI)	0.6039781	0.4796734	0.5006067	0.4564395	0.4134904	0.3295002	0.3428006	0.3123117	0.490903	0.3906516	0.4069405	0.3708648
	CTRL	0.5704747	0.4470749	0.4640672	0.4264579	0.4498083	0.3469828	0.3608475	0.3295447	0.503006	0.3907204	0.4059996	0.3717896
	TransformerXL	0.6371103	0.51427	0.5365341	0.4901452	0.4182796	0.3421597	0.3553643	0.324859	0.5050081	0.410921	0.4275488	0.3907417
	XLNet large	0.5815198	0.4671588	0.483462	0.4482108	0.4577767	0.3668583	0.3804321	0.350807	0.5122815	0.4109774	0.4258032	0.3935719
	DistilBERT base uncased	0.6130165	0.4918442	0.519021	0.472632	0.4375995	0.3591256	0.3780095	0.3451935	0.5106637	0.4151354	0.4374319	0.3989836
	ALBERT large v2	0.6278807	0.5132557	0.5197861	0.4894159	0.4081487	0.3301629	0.3366006	0.3129257	0.4947131	0.401836	0.4086012	0.3817596
	SciBERT uncased (allenAI)	0.5936238	0.4733964	0.4945263	0.4504344	0.4314072	0.3459681	0.3596508	0.3290732	0.4996797	0.3997734	0.4164401	0.3803065
	phoBERT large (vinAI)	0.5506821	0.4403838	0.4562341	0.4173047	0.446242	0.354763	0.368099	0.3363581	0.4929914	0.3929636	0.407455	0.3724844

Đánh giá

Kết quả tóm tắt theo từng chủ đề

		Rouge1	Rouge2	RougeL	RougeSU4	Rouge1	Rouge2	RougeL	RougeSU4	Rouge1	Rouge2	RougeL	RougeSU4
		p (Precision)				r (Recall)				f (Score)			
Giáo dục	BERT multilingual uncased	0.5362539	0.3681169	0.4118732	0.3499364	0.3574842	0.2452651	0.2732704	0.2306215	0.4289899	0.2943883	0.3285523	0.2780183
	BERT large uncased	0.5431566	0.354487	0.3996616	0.3325326	0.313515	0.207422	0.2287032	0.1935248	0.3975566	0.2617093	0.2909262	0.2446627
	GPT2	0.5843824	0.4114938	0.4544897	0.3891775	0.3051859	0.2106865	0.2320307	0.1958711	0.4009704	0.2786851	0.3072176	0.260589
	BART large (fb)	0.5643445	0.3789875	0.4058698	0.356827	0.3106367	0.2073465	0.2210834	0.1922667	0.4007083	0.2680443	0.2862449	0.2498879
	GPT (openAI)	0.5703636	0.3929506	0.4184749	0.3725365	0.3188322	0.2160945	0.2320559	0.2020273	0.4090219	0.2788446	0.2985548	0.2619815
	CTRL	0.4869314	0.3199976	0.3580515	0.3051143	0.3677203	0.2504663	0.2721225	0.236195	0.4190118	0.2809945	0.3092285	0.2662672
	TransformerXL	0.5835424	0.4060374	0.4562506	0.3791492	0.262237	0.1802198	0.1998312	0.1646615	0.3618589	0.2496378	0.2779321	0.2296067
	XLNet large	0.5258528	0.3523075	0.3915372	0.3310336	0.3524526	0.2440752	0.2666002	0.2272161	0.4220358	0.2883703	0.31721	0.2694714
	DistilBERT base uncased	0.5396576	0.354672	0.3979228	0.3330942	0.3154605	0.2083989	0.2304049	0.1928645	0.3981688	0.2625363	0.2918329	0.2442855
	ALBERT large v2	0.5116114	0.3319972	0.383816	0.3160148	0.3106544	0.2049765	0.2340625	0.1919113	0.3865766	0.2534635	0.2907916	0.2388017
	SciBERT uncased (allenAI)	0.5147101	0.3315891	0.3664473	0.3097318	0.3110579	0.1999457	0.2182652	0.1854682	0.3877715	0.2494656	0.2735796	0.2320089
	phoBERT large (vinAI)	0.4756506	0.2975964	0.3411803	0.2791251	0.3923966	0.2589271	0.2865914	0.2416608	0.4300312	0.2769183	0.3115124	0.2590454

Đánh giá

Kết quả tóm tắt theo từng chủ đề

		Rouge1	Rouge2	RougeL	RougeSU4	Rouge1	Rouge2	RougeL	RougeSU4	Rouge1	Rouge2	RougeL	RougeSU4
		p (Precision)				r (Recall)				f (Score)			
Kinh tế	BERT multilingual uncased	0.5303877	0.3565131	0.405619	0.333932	0.2997255	0.2004647	0.2262521	0.1859604	0.3830098	0.2566289	0.2904775	0.2388884
	BERT large uncased	0.4923362	0.3354714	0.3836286	0.3175698	0.2763583	0.186792	0.2126481	0.1747133	0.3540059	0.2399684	0.2736243	0.2254137
	GPT2	0.5690502	0.4095676	0.4505308	0.3888707	0.2804655	0.1989366	0.2190816	0.1856179	0.375741	0.2677976	0.2948064	0.2512891
	BART large (fb)	0.5184344	0.3402148	0.4003083	0.3249332	0.2860186	0.1909259	0.2174452	0.179552	0.3686527	0.2445899	0.2818119	0.2312948
	GPT (openAI)	0.5702345	0.4066653	0.4560027	0.3853901	0.3009573	0.2140852	0.2380279	0.1989542	0.3939803	0.2805025	0.3127855	0.2624308
	CTRL	0.5600802	0.4084399	0.4438196	0.3936602	0.3027009	0.2122832	0.2337101	0.2012951	0.3930006	0.2793675	0.3061862	0.2663792
	TransformerXL	0.5290678	0.3518177	0.4174424	0.3381706	0.2520871	0.1738697	0.2000281	0.1640864	0.3414718	0.2327255	0.270459	0.2209594
	XLNet large	0.5577696	0.3947037	0.4397571	0.3801336	0.3249828	0.2286561	0.2530737	0.2172619	0.4106826	0.2895644	0.3212645	0.2764953
	DistilBERT base uncased	0.5538531	0.3888964	0.4410771	0.3721555	0.3048118	0.2133949	0.2413219	0.201049	0.3932173	0.2755759	0.3119628	0.2610639
	ALBERT large v2	0.4945223	0.3219138	0.3816006	0.3100722	0.2607612	0.1715803	0.2002169	0.1616036	0.3414671	0.2238489	0.2626352	0.2124713
	SciBERT uncased (allenAI)	0.546434	0.3936239	0.4321241	0.37129	0.3048879	0.2142582	0.234931	0.1994181	0.391394	0.2774786	0.3043807	0.259474
	phoBERT large (vinAI)	0.5745262	0.4263386	0.4595856	0.4081024	0.3272755	0.232051	0.2536094	0.2180653	0.417006	0.3005281	0.3268538	0.2842464

Đánh giá

Kết quả tóm tắt theo từng chủ đề

		Rouge1	Rouge2	RougeL	RougeSU4	Rouge1	Rouge2	RougeL	RougeSU4	Rouge1	Rouge2	RougeL	RougeSU4
		p (Precision)				r (Recall)				f (Score)			
Văn hóa	BERT multilingual uncased	0.4199592	0.2606368	0.3184256	0.2467304	0.2516323	0.1601605	0.1890271	0.1513587	0.3147011	0.1984029	0.2372283	0.1876203
	BERT large uncased	0.4085676	0.2543534	0.3076702	0.2468952	0.208488	0.1262791	0.1526148	0.1205255	0.27609	0.1687692	0.2040259	0.1619788
	GPT2	0.4330206	0.2779414	0.3329663	0.2655267	0.2153727	0.1398192	0.1651709	0.1320788	0.2876674	0.1860469	0.220808	0.1764083
	BART large (fb)	0.4215892	0.270862	0.3287211	0.2607025	0.2250456	0.1507773	0.1759611	0.1434497	0.2934478	0.1937193	0.229222	0.1850673
	GPT (openAI)	0.4439052	0.2923267	0.3407382	0.2825137	0.2375897	0.1614728	0.1818087	0.1536302	0.3095175	0.2080337	0.2371047	0.199029
	CTRL	0.4073576	0.2533981	0.3041841	0.2413711	0.240232	0.1446529	0.1715878	0.1363432	0.3022295	0.1841712	0.2194089	0.174255
	TransformerXL	0.4385442	0.2668152	0.3352392	0.2556348	0.2181673	0.1381084	0.1656958	0.1312833	0.2913791	0.1820068	0.2217762	0.1734764
	XLNet large	0.3810868	0.2194177	0.2730563	0.2109415	0.2459152	0.1416254	0.1724529	0.1352564	0.2989306	0.1721408	0.2113957	0.1648259
	DistilBERT base uncased	0.4213926	0.2537214	0.3145166	0.2466508	0.2291694	0.1459588	0.1744636	0.1391974	0.2968826	0.1853125	0.2244332	0.177962
	ALBERT large v2	0.4184157	0.2593092	0.3127455	0.2433736	0.2397111	0.1548055	0.1787942	0.1444278	0.3048011	0.1938714	0.2275181	0.181278
	SciBERT uncased (allenAI)	0.4097154	0.2510904	0.3092636	0.2391143	0.2373394	0.1514227	0.1799439	0.1441804	0.3005669	0.188917	0.2275112	0.1798908
	phoBERT large (vinAI)	0.4246694	0.2655934	0.323493	0.2493474	0.2602404	0.1706796	0.1984433	0.160637	0.3227174	0.207812	0.2459879	0.1953948

Đánh giá

Kết quả tóm tắt theo từng chủ đề

		Rouge1	Rouge2	RougeL	RougeSU4	Rouge1	Rouge2	RougeL	RougeSU4	Rouge1	Rouge2	RougeL	RougeSU4
		p (Precision)				r (Recall)				f (Score)			
Xã hội	BERT multilingual uncased	0.5007175	0.3358911	0.3770743	0.3190642	0.3041657	0.2110283	0.2315935	0.1987763	0.3784427	0.2592065	0.2869479	0.2449496
	BERT large uncased	0.5034775	0.3443999	0.3808208	0.3289225	0.2582426	0.1768757	0.1954962	0.165626	0.3413835	0.2337189	0.2583614	0.2203145
	GPT2	0.5406925	0.3802987	0.4236413	0.365884	0.2674637	0.1892316	0.2090204	0.1786077	0.3578903	0.2527154	0.2799274	0.2400393
	BART large (fb)	0.5317806	0.3642061	0.4169197	0.3507349	0.2943346	0.207849	0.2306529	0.1974944	0.3789336	0.2646594	0.2969976	0.2526978
	GPT (openAI)	0.4878402	0.3171096	0.3700457	0.306068	0.2677151	0.1753157	0.201075	0.1656045	0.3457118	0.2257979	0.2605647	0.2149213
	CTRL	0.5074086	0.3569204	0.3761796	0.3416211	0.3050617	0.2115515	0.2233297	0.1992033	0.3810378	0.2656491	0.2802695	0.2516604
	TransformerXL	0.525901	0.3523827	0.3980211	0.3347149	0.245627	0.1682834	0.1868989	0.1566072	0.3348563	0.2277858	0.2543586	0.2133784
	XLNet large	0.5192723	0.3745681	0.4022065	0.3535304	0.2946413	0.215823	0.2293807	0.2007262	0.3759589	0.2738538	0.2921478	0.2560648
	DistilBERT base uncased	0.5045493	0.3292222	0.3768332	0.3167572	0.2881427	0.1959244	0.2159299	0.185842	0.3668062	0.2456558	0.2745432	0.2342495
	ALBERT large v2	0.4805434	0.3075153	0.3545977	0.2925807	0.2568887	0.1653194	0.1907307	0.1551245	0.3348001	0.215036	0.2480438	0.2027514
	SciBERT uncased (allenAI)	0.4852294	0.3236983	0.363979	0.3051053	0.2689492	0.1786521	0.2007153	0.1658673	0.3460773	0.2302352	0.2587459	0.2149042
	phoBERT large (vinAI)	0.4869446	0.3202473	0.3641658	0.3029999	0.3074517	0.2086409	0.230677	0.194424	0.3769201	0.2526685	0.2824432	0.2368622



Đánh giá kết quả

Dựa trên số liệu của dataset này, ta thấy BERT và các model cải tiến của nó (phoBERT, XLNet) cho kết quả lý tưởng hơn so với các model khác của Transformers.

Bên cạnh đó, kết quả tóm tắt có sự khác nhau giữa các chủ đề. Cụ thể với chủ đề **Công nghệ** cho kết quả tốt nhất, **Văn hóa** cho kết quả thấp nhất.

Có thể kết luận, **đặc trưng về chủ đề** cũng là một đặc trưng quan trọng ảnh hưởng mạnh đến độ chính xác của bài toán tóm tắt văn bản.



Đánh giá kết quả

Ngoài ra, kết quả tóm tắt chưa cao còn do những nguyên nhân sau:

- **Data đầu vào:** chưa sạch, không đồng bộ về encoding (mã hoá) giữa văn bản tóm tắt hệ thống và tóm tắt mẫu. Văn bản gốc VH02, VH03, VH04, văn bản mẫu CT17, VH11 encoding là “ucs-2 le bom”, thay vì “utf-8”. Các bảng kết quả trên chạy với đầu vào đã được xử lý thủ công, với các bộ dataset khác có thể gặp lỗi gây nhiễu số liệu, cần có hướng giải quyết triệt để.

- Yếu tố con người

- + **Tóm tắt mẫu không khớp với văn bản gốc.** Cụ thể văn bản “KT15.TXT”, nội dung gốc liên quan đến kinh tế, nhưng tóm tắt mẫu nội dung lại nói về giáo dục.
- + **Độ dài giữa tóm tắt mẫu và tóm tắt của model.** Không có chuẩn chung cố định cho độ nén (Ratio) hay số câu (Number of sentences) trong văn bản tóm tắt.

Kết luận

Model có những ưu điểm như:

- Không đòi hỏi domain knowledge quá nhiều. Người không chuyên sâu về ngôn ngữ học vẫn có thể xây dựng được ứng dụng tóm tắt văn bản.
- Thuật toán rõ ràng, dễ nắm bắt các tham số, có lợi thế khi tích hợp thêm tri thức.
- Tận dụng được các thư viện về ngôn ngữ và pretrained model, không cần trực tiếp training (*vẫn có thể train thêm để tăng hiệu quả*), thích hợp với những ngôn ngữ ít tài nguyên (bộ dữ liệu chuẩn) như tiếng Việt.
- Có thể tóm tắt các văn bản lớn. Đây là ưu điểm so với tóm tắt tóm lược (Abstraction), các mô hình Abstraction như Seq2seq gặp khá nhiều khó khăn trong việc tóm tắt văn bản lớn.

Kết quả bước đầu cho thấy hiệu quả khá tốt.

Việc cần làm trong tương lai

- Xử lý các vấn đề tồn đọng trong model hiện tại như làm sạch data đầu vào, chuẩn hóa data mẫu, ...
- Giải quyết mối quan hệ giữa chủ đề văn bản với các bản tóm tắt, phân tích và tìm thuật toán phù hợp để bản tóm tắt cô đọng và xoáy sâu và chủ đề hơn.
- Thu thập thêm data mẫu để hỗ trợ việc đánh giá.
- Nghiên cứu, áp dụng các phương pháp mới nâng cao chất lượng văn bản tóm tắt. Dựa trên cơ sở kiến thức đã tìm hiểu, có thể tiến hành xây dựng hệ thống tóm tắt theo tóm lược (Abstraction) hoặc mở rộng loại hình nghiên cứu (tóm tắt nội dung từ các đầu vào đa phương tiện như hình ảnh, âm thanh, video, ...)

Tài liệu tham khảo

[1] Khóa học Deep Learning cơ bản - ThS.Nguyễn Thanh Tuấn, founder AI4E

[2] Dataset tiếng Việt - PGS.TS.Lê Thanh Hương, ĐH Bách khoa Hà Nội

[3] Thư viện Transformers - Hugging Face (huggingface.co)

[4] Thư viện ngôn ngữ spaCy và module Neural Coreference (spacy.io)

[5] "Deep Reinforcement Learning for Mention-Ranking Coreference Models" - EMNLP 2016

"Improving Coreference Resolution by Learning Entity-Level Distributed Representations"
- ACL 2016

by Kevin Clark and Christopher D. Manning

[6] Tham khảo tính năng các model BERT, DistilBERT, XLNet, ... (trituenhantao.io, ichi.pro)

[7] Paper BERT for Extractive Text Summarization on Lectures - Derek Miller (arxiv.org/pdf/1906.04165)

[7] "ROUGE: a Package for Automatic Evaluation of Summaries", In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), Barcelona, Spain, July 25 - 26, 2004 - Lin, Chin-Yew

[8] Giải quyết vấn đề (stackoverflow.com, ...)



Thanks & Best Regards!

Chanh Vo (Mr. Cyber)

Tel:

+84 984 320 841

E-mail:

chanhvokts@gmail.com