**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC SƯ PHẠM HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**



**TEXT SUMMARIZATION**

**GVHD:**

**Nguyễn Hồng Bửu Long**

**Sinh viên:**

**Hồ Khả Việt Huấn – 43.01.104.058**

**Nguyễn Văn Thịnh – 43.01.104.168**

**Lâm Phước Đạt – 43.01.104.015**

**THÀNH PHỐ HỒ CHÍ MINH**

**Mục Lục**

[**I/ Giới thiệu xử lý ngôn ngữ tự nhiên:** 3](#_Toc57936584)

[1) Khái niệm 3](#_Toc57936585)

[2) Ứng dụng 3](#_Toc57936586)

[**II/ Hệ thống tóm tắt văn bản tự động (Text summarization** 3](#_Toc57936587)

[1. Giới thiệu đề tài 3](#_Toc57936588)

[1.1. Khái Niệm 3](#_Toc57936589)

[1.2. Lí do 3](#_Toc57936590)

[1.3. Ứng dụng 3](#_Toc57936591)

[**2.** **Phân Loại** 3](#_Toc57936592)

[2.1. Tóm tắt trích xuất (extractive summarization) 3](#_Toc57936593)

[2.2. Tóm tắt trừu tượng (abstractive summarization) 4](#_Toc57936594)

[2.3. Tóm tắt hỗn hợp (Hybrid) 4](#_Toc57936595)

[**3.** **Giải Pháp** 4](#_Toc57936596)

[3.1 Giới thiệu CNN 4](#_Toc57936597)

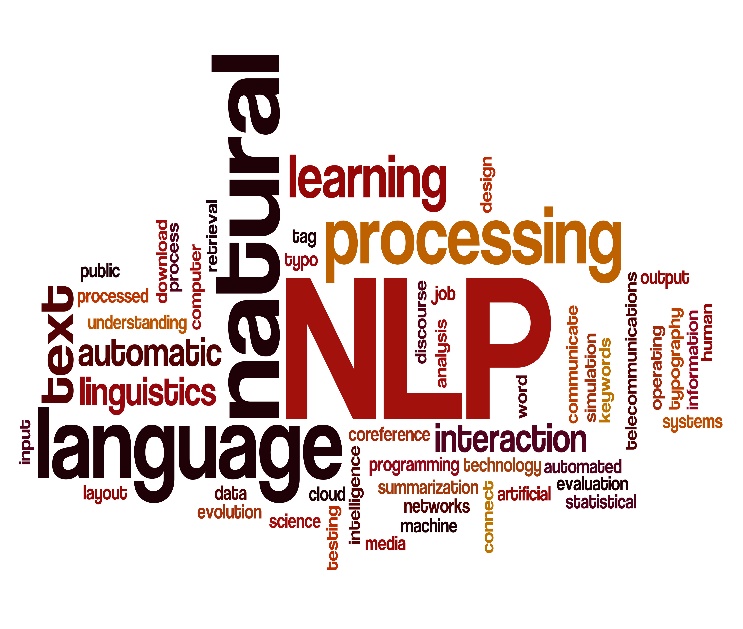
[3.2 Giới thiệu về Word2vec 5](#_Toc57936598)

[3.3 Kết quả 5](#_Toc57936599)

[3.4 Hướng phát triển 5](#_Toc57936600)

[**III/ Tài liệu tham khảo** 5](#_Toc57936601)

# **I/ Giới thiệu xử lý ngôn ngữ tự nhiên:**

1. **Khái niệm**

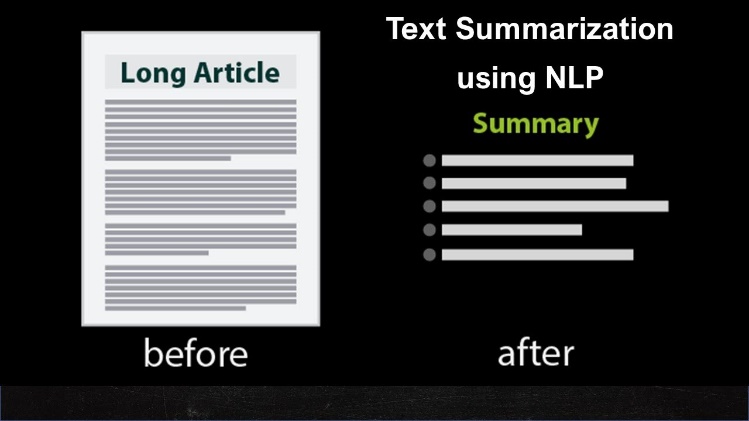
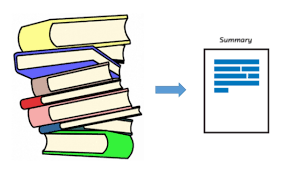
* Xử lý ngôn ngữ tự nhiên (Natural Language Processing – NLP) là một trong những nhánh khó của trí tuệ nhân tạo. Bởi lẽ ngôn ngữ là một hệ thống phức tạp để giao tiếp giữa những động vật bậc cao hay có năng lực tư duy như con người. Nếu NLP được giải quyết thành công đồng nghĩa với việc máy tính có thể hiểu và sử dụng ngôn ngữ tự nhiên để giao tiếp như chúng ta.

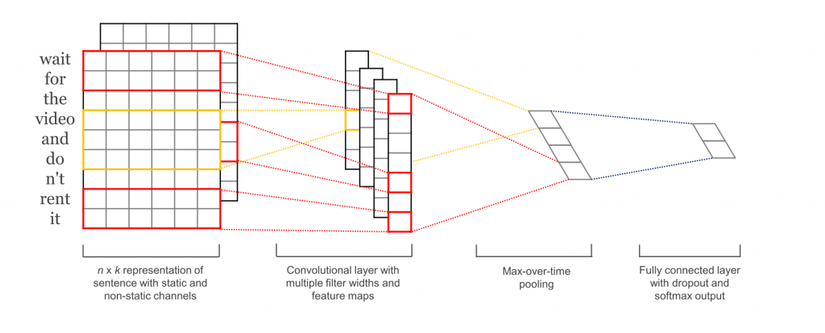
1. **Ứng dụng**

* Machine translation
* Information Retrival
* Sentiment Analysis
* Information Extraction
* Question Answering

# **II/ Hệ thống tóm tắt văn bản tự động (Text summarization****)**

1. **Giới thiệu đề tài**
   1. ***Khái Niệm***

* Tóm tắt văn bản tự động nhằm mục đích trích xuất nội dung từ nguồn thông tin và trình bày các nội dung quan trọng nhất cho người sử dụng theo một khuông dạng súc tích và gây cảm xúc đối với người sử dụng một chương trình cần đến.
  1. ***Lí do***
  2. ***Ứng dụng***
* Tóm tắt tin tức



* Tóm tắt cảm xúc
* Tóm tắt tiểu thuyết, truyện tranh
* Tóm tắt các tài liệu pháp lý

1. **Phân Loại**
   1. ***Tóm tắt trích xuất (extractive summarization)***
2. *Định nghĩa*

* Đánh giá mức độ quan trọng của một số câu hoặc từ quan trọng trong đoạn văn bản gốc và sau đó chọn ra những câu hoặc từ có điểm cao nhất gom lại với nhau tạo thành một bản tóm tắt.

1. *Các phương pháp*

* Machine learning
* Deep-learning
* Optimazation
* Fuzzy logic
  1. ***Tóm tắt trừu tượng (abstractive summarization)***

1. *Định nghĩa*

* Sử dụng kỹ thuật NLP nâng cao để sinh ra một đoạn văn bản tóm tắt mới. Trong đoạn văn bản tóm tắt có thể không có những từ ở trong văn bản gốc.

1. *Các phương pháp*

* Deep-learning
* Structure-based
* Sementic-based
  1. ***Tóm tắt hỗn hợp (Hybrid)***

1. *Định nghĩa*

* Mô hình này là sự kết hợp giữa tóm tắt theo trích xuất và tóm tắt trừu tượng

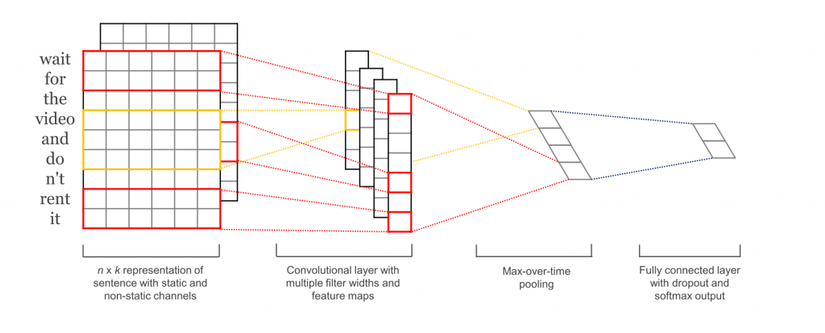
1. *Các phương pháp*

* Extractive to Abstractive
* Extractive to Shallow Abstractive

1. **Giải Pháp**
   1. ***Giới thiệu CNN***
2. *Khái niệm CNN*

* Convolutional Neural Network(CNN) là một trong những mô hình phổ biến của Deep Learning. CNN được sử dụng trong nhiều bài toán như nhận dạng ảnh, phân tích video hoặc các bài toán về xử lý ngôn ngữ tự nhiên

1. *CNN cho xử lý ngôn ngữ tự nhiên*

* Sau khi chạy Word2vec thì cho ra từ biểu diễn dưới dạng vector, từ đó các vector kết nối với nhau thành một ma trận. Ma trận lúc này biểu diễn cho một câu.
* Từ đây chúng ta có thể xem từng hàng là từng từ trong câu. Chúng ta sử dụng CNN để học được các cụm từ trong câu. Các lớp CNN tạo ra những feature mới từ những cụm từ đó.
  1. ***Giới thiệu về Word2vec***
* Word2vec được tạo ra từ năm 2013 nó là một mô hình unsupervised learning. Số chiều của Word2vec là NxD với N số lượng văn bản và D là số chiều word embedding. Word2vec có 2 mô hình chính là skip-gram và Cbow:
  + Skip-gram là mô hình dự đoán từ xung quanh khi cho một từ cho trước. Nhiêm vụ của mô hình này là khi nhập vào một từ nào đó, thì mô hình sẽ cho ra những từ liên quan đến từ đã nhập.
  + Cbow là viết tắt của Continous bag of word. Mô hình này ngược lại với mô hình skip-gram là cho những từ xung quanh và dự đoán từ liên quan.
  1. ***Kết quả***
  2. ***Hướng phát triển***

# **III/ Tài liệu tham khảo :**

1. <https://ieeexplore.ieee.org/document/7508049>
2. <https://www.analyticsvidhya.com/blog/2019/06/comprehensive-guide-text-summarization-using-deep-learning-python/>