



ĐẠI HỌC BÁCH KHOA HÀ NỘI  
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

# Natural Language Processing group

Presenter: Le Thanh Huong  
School of Information and Communication Technology  
Hanoi University of Science and Technology

# Members

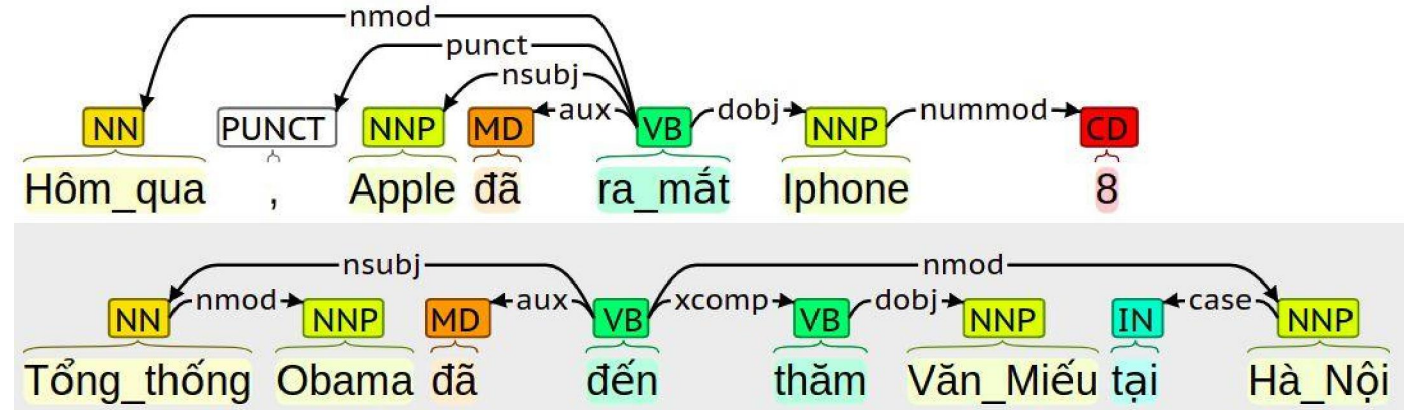
- Le Thanh Huong
- Nguyen Kiem Hieu
- Nguyen Thi Thu Trang
- Tran Viet Trung
- Cao Tuan Dung
- Nguyen Thi Kim Anh
  
- + students

# BK.AI - NLP group - Research directions

- Vietnamese natural language tools: Pyvi, vi\_spacy (Trung), BK Treebank (Hieu)
- Information extraction (named entity recognition, relation extraction, event extraction) (Hieu, Huong, Trung)
- Weak supervised taxonomy construction (Trung, Dung) – Naver project
- Knowledge base construction from semi-structured documents (Trung)
- Vietnamese spelling correction (Huong)
- Plagiarism checking (Huong, Trang)
- Text summarization (Huong, Trang, Kim Anh)
- Machine translation (Trung, Huong)
- Chatbot, question answering (Trung, Huong, Trang, Hieu, Dung)
- Sentiment analysis (Hieu, Huong)

# BKTreebank

A Vietnamese dependency parser



## Diacritic restoration

- Input: “*Hôm nay tôi đi học*”
- Output: “*Hôm nay tôi đi học*” (Today I go to school)
- Training data are freely available
- Approaches: Sequence labelling/seq-to-seq
- Challenges: Subtle ambiguity requires high-level syntactic/semantic understanding

Kiem-Hieu Nguyen, Cheol-Young Ock. *Diacritics restoration in Vietnamese: letter based vs. syllable based model*. PRICAI 2010

# Part-of-speech tagging Named-Entity Recognition

- biLSTM + Conditional Random Fields
- Utilize char-level representation

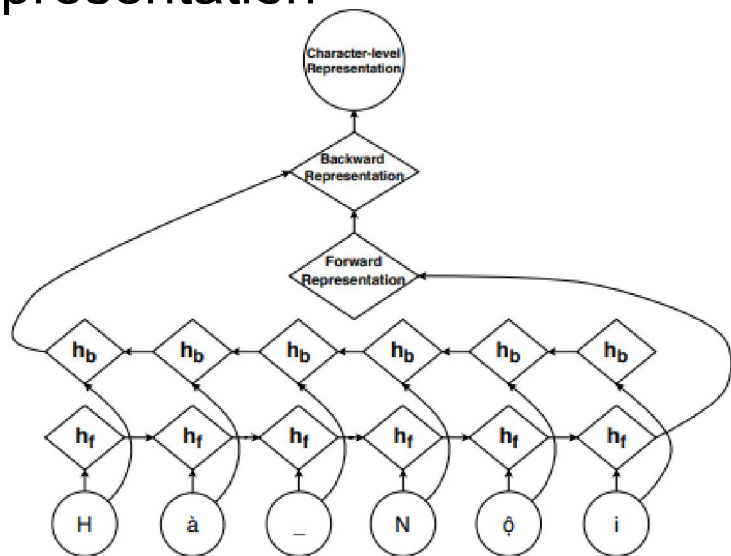
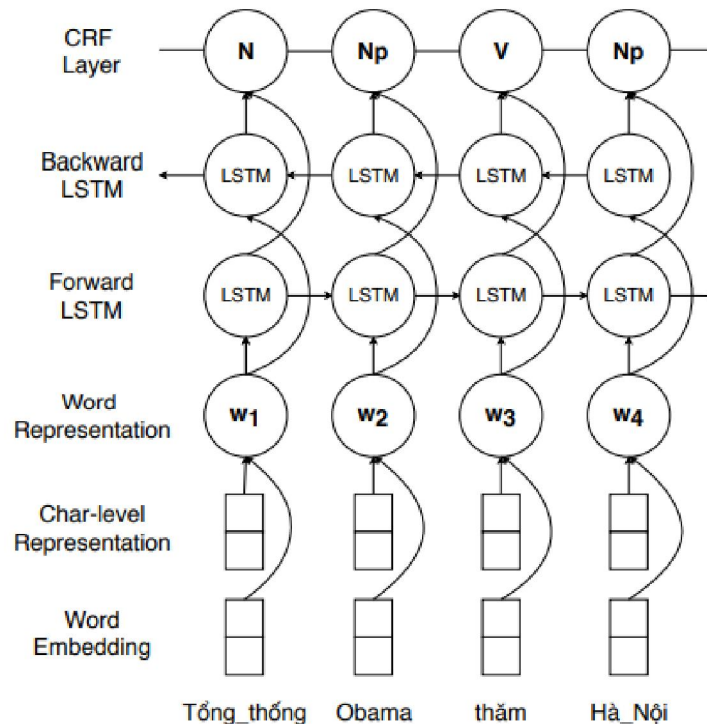


Figure 2. The character embeddings of the word “Hà\_Nội” are given to a bidirectional LSTMs for extracting character-level word features. Character-level representation of one word is the concatenation of its forward and backward representations



# Knowledge base construction from semi-structured documents

## Easily capture applicant data with our Resume AI

Making your HR software smarter, save time on manual data entry. Our Resume AI engine uses advanced Artificial Intelligence and Machine Learning techniques to automatically extract every CV into a complete and searchable candidate record.



**Lê Thanh Tuấn**

Ngày sinh: 17/05/1998  
Giới tính: Nam  
Điện thoại: [protected data]  
Email: [protected@parser]  
Địa chỉ: Phố Tương Mai - Đông Tâm - Hai Bà Trưng - Hà Nội

**MỤC TIÊU NGHỀ NGHIỆP**

Mục tiêu trước mắt là mong muốn được tham gia học hỏi thêm kiến thức mới, tích lũy kinh nghiệm. Áp dụng kỹ năng, kiến thức của bản thân để tham gia cùng xây dựng sản phẩm. Cùng kinh nghiệm có được trong quá trình học hỏi, với niềm đam mê sẵn có, luôn cố gắng để có thể góp phần tạo nên được sản phẩm tốt nhất.

**HỌC VẤN**

09/2016-nay **Đại Học Bách Khoa Hà Nội**  
Hiện tại đang sinh viên năm 3 chuyên ngành Công nghệ thông tin - Điểm tổng CPA đến nay 3.11

**CÁC KỸ NĂNG**

**Kỹ năng lập trình**

- Có nền tảng cơ bản về Java, JavaScript, Python
- Kiến thức cơ bản về Machine Learning, Deep Learning
- Kiến thức cơ bản về xử lý ngôn ngữ tự nhiên
- Biết sử dụng cơ bản hệ quản trị cơ sở dữ liệu MySQL, MongoDB
- Sử dụng tốt các IDE như là Eclipse, PyCharm, VisualCode, JupyterNotebook
- Biết sử dụng Git

**Các kĩ năng khác**

- Có khả năng làm việc nhóm tốt
- Có khả năng sắp xếp thu tốt
- Có thái độ tích cực trong công việc
- Có khả năng tự tìm hiểu những kiến thức, công nghệ mới

**KINH NGHIỆM**

- Đã làm bài tập lớn về Web với Nodejs, sử dụng Express Framework, MongoDB
- Đã làm bài tập lớn về hệ thống quản lý đào tạo sử dụng Java Spring, MySQL
- Đã làm bài tập lớn về xử lý ngôn ngữ tự nhiên, phân tích cảm xúc bình luận

**NGOẠI NGỮ**

- Có khả năng đọc hiểu tốt các bài viết bằng tiếng Anh



# Knowledge base construction from semi-structured documents



Tìm kiếm



Trang chủ

Khu vực

Lĩnh vực

Vị trí công việc

Công ty

Đăng nhập

Đăng ký

## Thông tin phân tích thị trường lao động Việt Nam

Công ty ▾

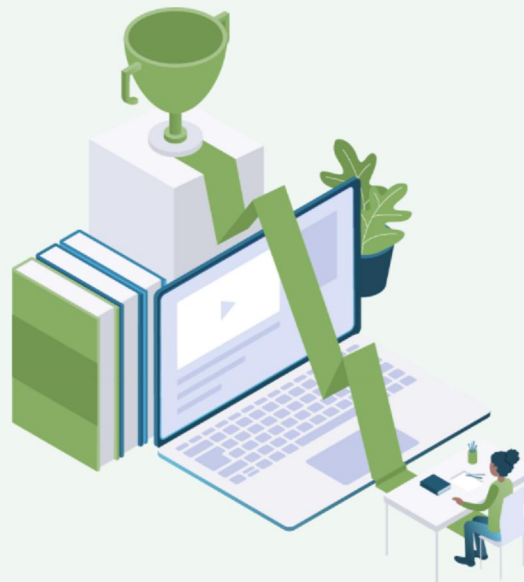


15000+

Công ty

120000+

Công việc



### Phân tích khu vực

Phân tích nhân khẩu học về các góc cạnh khác nhau của thị trường lao động như tính vùng miền, nghề nghiệp, kỹ năng, độ tuổi, giới tính, nền tảng giáo dục, chế độ đãi ngộ

Khám phá >



### Phân tích theo lĩnh vực

Phân tích nhân khẩu học về các góc cạnh khác nhau của thị trường lao động như tính vùng miền, nghề nghiệp, kỹ năng, độ tuổi, giới tính, nền tảng giáo dục, chế độ đãi ngộ

Khám phá >



### Phân tích theo vị trí

Phân tích phân bố việc làm theo ngành nghề, vị trí địa lí. Dự báo nhu cầu của các vị trí công việc trong tương lai. Tìm kiếm các nhóm kỹ năng phù hợp với vị trí công việc cụ thể.



### Phân tích công ty

Phân tích mức lương, chế độ đãi ngộ cho các vị trí tuyển dụng của từng công ty cụ thể. Thống kê các vị trí tuyển dụng của công ty theo thời gian, theo khu vực vùng miền

# Plagiarism checking



HỆ THỐNG QUẢN TRỊ ĐẠI HỌC TRỰC TUYẾN  
TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI - VIỆN CÔNG NGHỆ THÔNG TIN VÀ  
TRUYỀN THÔNG



Tuần 37 (Tuần học thứ 13): [17/05 - 22/05](#)

TRANG CHỦ   GIẢNG DẠY   DANH SÁCH ĐỒ ÁN   KHOA HỌC CN   CSVC   CÁN BỘ   SINH VIÊN   Vai trò: Giảng viên

Bộ lọc: 20202   Đồ án Môn học   Loại ĐA   Học phần   Trạng thái   MSSV, Họ tên SV ...

Đề án tốt nghiệp: 14 ĐA

1-14 of 14   Số dòng: 100

#	Mã HP	Mã Lớp	MSSV	Hệ	Tên học phần	Đơn vị	GVHD	Báo cáo	Trùng	G	
<input type="checkbox"/>	1	IT4995	705373	2017312	CNKT	<a href="#">Đề án tốt nghiệp cử nhân</a> <a href="#">Các đề tài khác do sinh viên</a> <a href="#">tự đề xuất</a>	HTTT	HTTT - Lê Thanh Hương	<input type="checkbox"/>	<a href="#">N/A</a>	1.

Select All

Kiểm tra độ Trùng lặp với Nguồn các tài liệu khác

Kiểm tra Trùng lặp





# Plagiarism checking

- The system can check plagiarism for theses in Vietnamese
- Database: Theses from different universities; journal papers in Vietnamese
- Checking:
  - Compare with documents in the system, do not check with online documents
  - Checking sentential similarity based on an improvement of Levenshtein distance measure, concerning word reordering, distance between words, ...
  - Exclude short sentences (<7 words), acknowledgment, TOC
- Future directions:
  - Challenge: Documents from Internet
  - Implementing a function for checking documents in English
  - Developing a flexible interface: people can accept/reject the highlights of similar text, exclude sources, add comments
  - Developing a function for spell checking

*Nguyen Van Son, Le Thanh Huong, Nguyen Chi Thanh. A two-phase plagiarism detection system based on multi-layer LSTM Networks. IAES International Journal of Artificial Intelligence, 2021*

# Coopy (sentence-based)

là bu 20121323\_Tran\_Van\_Chung\_1496405173700.pdf

dùng 2.3.2 Mô hình không gian vector Mô hình không gian vector là một mô hình đại số thể hiện thông tin văn  
gian bản như một vector, các phần tử của vector này thể hiện mức độ quan trọng của một từ và cả sự xuất hiện  
hay không xuất hiện (Bag of words) của nó trong một tài liệu

## 3.7.1 Vector Space Model (VSM)

Vector space model là một mô hình đại số (algebraic model) biểu diễn thông tin văn bản như một vector, các phần tử của vector này thể hiện mức độ quan trọng của một từ và cả sự xuất hiện hay không xuất hiện của nó trong một tài liệu.

# TurnItIn (word-based)

< "ECAI 2020", IOS Press, 2020 >

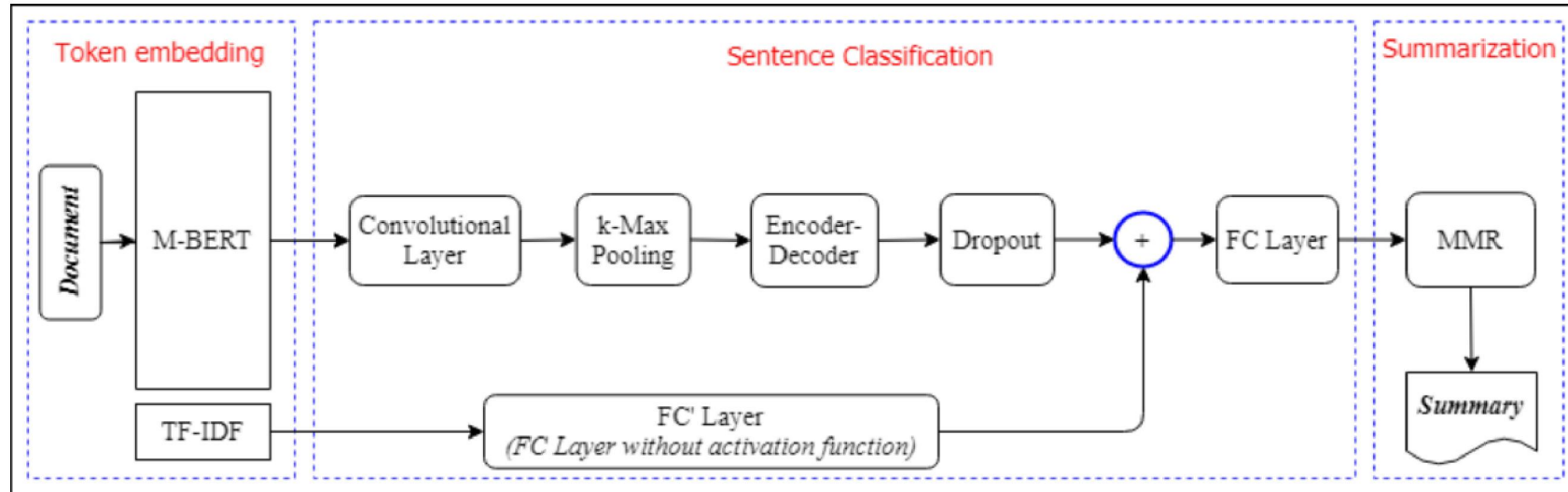
Publication

beddings. Finally, we study how to improve the performance of a model considering graphs coming from related datasets into the training procedure in a semi-supervised learning fashion. 1 INTRODUCTION State-of-the-art machine learning techniques for classification and regression on graphs are at the moment kernel machines equipped with specifically designed kernels for graphs (e.g, [31, 36]). Although there are examples of kernels for

s of social sites) is unstruc-  
texts combined with metadata  
numbers, username, id, etc).  
partially missing and incorrect,  
: data more challenging.  
odal learning, which 'involves  
multiple sources' [20] result-

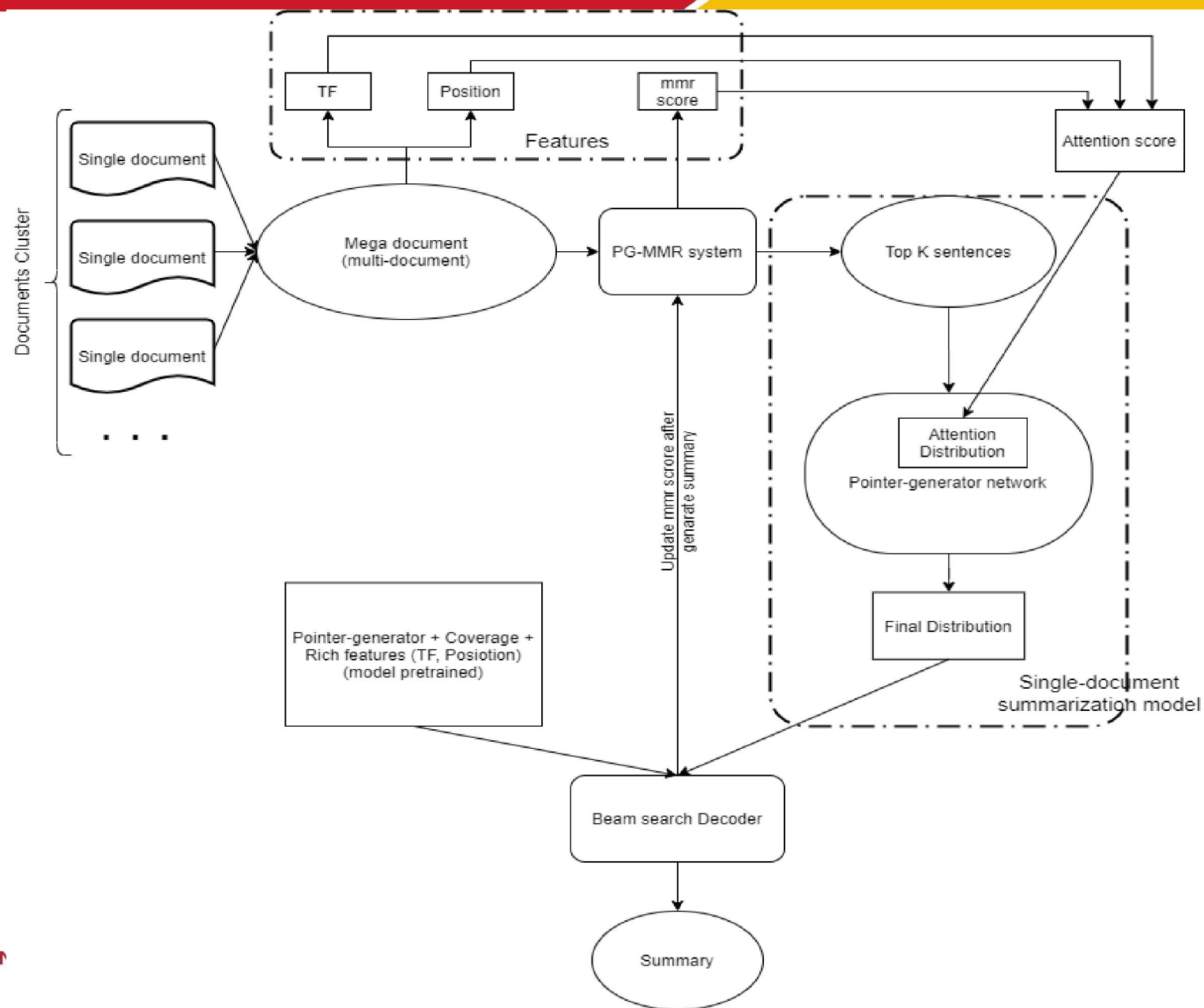
- light on the meta data's informativeness.
- We apply state-of-the-art transfer learning methods for textual feature extractions and neural network (in comparison with other traditional machine learning methods) for tabular-data feature representation, achieving the competitive performance of 0.9418 ROC-score on the public test set (ranked 2nd) and 0.9467 ROC-score (ranked 3th)

# Single, extractive summarization



1. Minh-Tuan Luu, Thanh-Huong Le, Minh-Tan Hoang (2021). An effective deep learning approach for extractive text summarization. Indian Journal of Computer Science and Engineering. Vol. 12, No. 2, 2021, e-ISSN:0976-5166, p-ISSN:2231-3850, pp.434-444.
2. Minh-Tuan Luu, Thanh-Huong Le, Minh-Tan Hoang . A hybrid model using the pretrained bert and deep neural networks with rich feature for extractive text summarization. Journal of Computer Science and Cybernetics, 2021 (accepted)

# Multiple, abstractive summarization

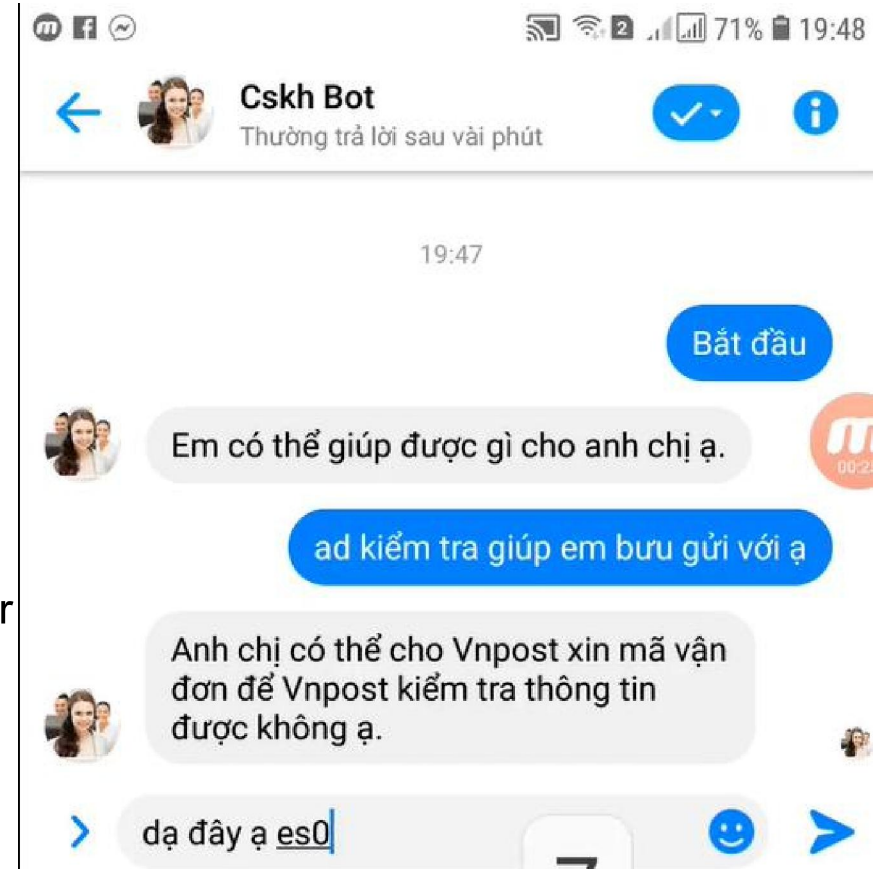


# Multilingual Machine Translation

- Cooperate with VNU University of Engineering and Technology
- Basing on neural machine translation – Open NMT
- Creating sentence alignment datasets (Vietnamese – Khmer):
  - Text alignment
    - Translate Khmer documents into Vietnamese using Google translation API
    - Extracting a bag of words for each document
    - Compute cosine similarity based on 2 bags of words
  - Sentence alignment
    - Translate Khmer sentences into Vietnamese using Google translation API
    - Compute sentential similarity: using Bert for sentence embedding

# Chatbot – QA system

- With an available QA set:
  - Measure question similarity based on question types, cosine similarity, BERT
- Without any QA dataset
  - Question classification: SVM, CNN
  - Named entity recognition: NeuroNER, FlexCRF
  - Searching for results
  - Generating answers based on predefined templates or summarizing answers
- Future work:
  - Use pretrain model like BERT, Electra, fine tune with QA datasets, in order to generate answers from the input questions

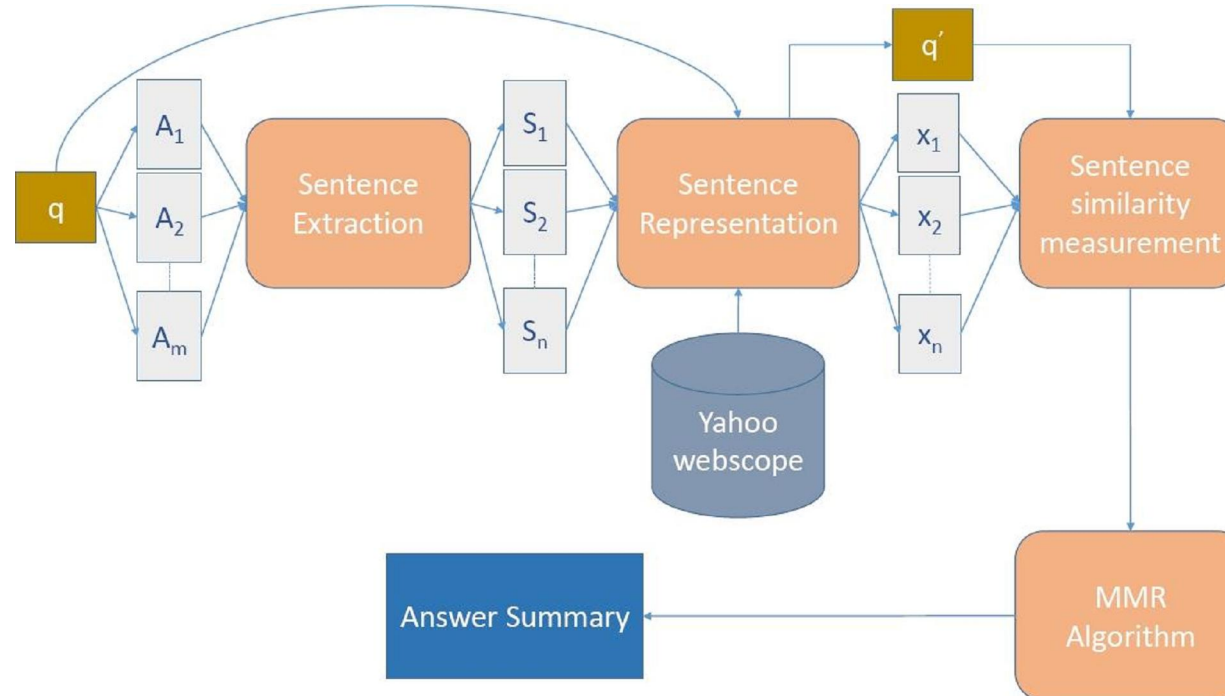


Huong T. Le, Dung T. Cao, Trung H. Bui, Long T. Luong and Huy Q. Nguyen (2021). Improve Quora Question Pair Dataset for Question Similarity Task. RIVF 2021.



# CQA: Answer summarization

- Sentence representation from auto-encoder
- Maximal Marginal Relevance framework



Thi-Thanh Ha, Thanh-Chinh Nguyen, Kiem-Hieu Nguyen, Van-Chung Vu, Kim-Anh Nguyen. *Unsupervised Sentence Embeddings for Answer Summarization in Non-factoid CQA*. CICLING 2018

# CQA: Vietnamese question retrieval

- A dataset for question retrieval in Vietnamese in the mobile domain
- A BERT model for the domain

	<b>Pairs of questions</b>
Train	5,996
Dev	847
Test	1,068
Average length (syllable)	27
Vocabulary (syllable)	5,821

Table 1: Statistics of Thegioididong dataset.

Corpus size	1.1M
Vocabulary size (syllable)	151,735
Average length (syllable)	31

Table 2: Statistics of unlabeled corpus crawled from The gioi Di dong.

<b>Models</b>	<b>MAP</b>
LSTM	52.60
CNN	53.10
ABCNN	51.52
LSTM attention	55.50
BERT-multilingual	61.06
BERT4Vn	63.75
PhoBERT	65.50
BERT4ecommerce	<b>70.50</b>
ElasticSearch	52,00
SVM	49.75

Thi-Thanh Ha, Van-Nha Nguyen, Kiem-Hieu Nguyen, Kim Anh Nguyen, Tien-Thanh Nguyen. *Utilizing Bert for Question Retrieval on Vietnamese E-commerce Sites*. PACLIC 2020

Thank you for your attention!

Q & A