Vietnamese Text Retrieval : Test Collection and First Experimentations

Ho Bao Quoc

Vietnam National University

HoChiMinh City University of Sciences

Where are we ?


© MAPQUEST.
0 1,000 Miles
0 1,000 Kilometers
60° N
40° N
Asia
Bering Sea
Tropic of Cancer
180°
20° N
EUROPE
Georgia
Armenia
Ural
Russia
Turkey
Azerbaijan
Kazakhstan
Cyprus
Lebanon
Israel
Syria
Jordan
Iraq
Uzbek.
Turkmen.
Mongolia
North Korea
Japan
Kyrgyzstan
Tajikistan
Iran
Afghan.
China
South Korea
PACIFIC OCEAN
Kuwait
Bahrain
Qatar
Saudi Arabia
U.A.E.
Pakistan
Bhutan
Nepal
160° E
AFRICA
Yemen
Oman
India
Hong Kong (China)
Laos
Macao (China)
Philippines
Arabian Sea
Myanmar
Vietnam
Cambodia
Philippine Sea
Bay of Bengal
Brunei
Malaysia
Maldives
Sri Lanka
Indonesia
Singapore
Timor-Leste
INDIAN OCEAN
AUSTRALIA
0°
20° S
40° E
60° E
80° E
100° E
120° E

I am here !!!


Vietnam
MAPQUEST
CHINA
MYANMAR (BURMA)
LAOS
Indochina
THAILAND
CAMBODIA
Gulf of Tonkin
Hainan
South China Sea
Andaman Sea
Gulf of Thailand
Isthmus of Kra
Phu Quoc I.
Paracel Islands
(Administered by China—
Claimed by Vietnam)
Palawan
(PHILIPPINES)
Con Dao
Annam Cordillera
Mekong
Red
Ping
Phan Si Pan
10,311 ft.
(3,143 m)
Lao Cai
Cao Bang
Viet Tri
Thai Nguyen
Son La
Hong Gai
Hanoi
Haiphong
Nam Dinh
Thanh Hoa
Vinh
Dong Hoi
Dong Ha
Hue
Da Nang
Quang Ngai
Kon Tum
Play Cu
Quy Nhon
Tuy Hoa
Buon Ma Thuot
Nha Trang
Da Lat
Phan Rang-Thap Cham
Bien Hoa
Ho Chi Minh City
Phan Thiet
Vung Tau
Long Xuyen
My Tho
Can Tho
0 120 Miles
0 120 Kilometers
20° N
16° N
12° N
8° N
94° E
98° E
102° E
106° E
110° E
114° E

- Faculty of Information Technology

  HoChiMinh City University of Sciences

  Vietnam National University

  227 Nguyen Van Cu – 5 District – HoChiMinh City – Vietnam

  hbquoc@fit.hcmuns.edu.vn

Plan

- Vietnamese specialities
- Vietnamese Test Collection
- Experimentations

Vietnamese Specialities

Vietnamese Alphabet

- Monosyllabic language
- Latin based Alphabet with accents on vowels

  Ex: ă, â, ê, ô, ư
- Usage six tons : (bằng), ‘ (sắc), ` (huyền), ? (hỏi) ~ (ngã), . (nặng) : the word sense is changed with the different tons :

Tons example

Ex : ma = phantom

má = cheek

mà = but

mả = tomb

mã = code

mạ = rice seedling

=> There are many character-sets : ABC, TCVN, VNI, UFT-8.

Vietnamese word

- Linguistic unit : “tiếng” : string of characters separated with another by one white bank
- Word contain one or more “tiếng”

  Ex. Sách = book

  dữ liệu = data

  xã hội chủ nghĩa = socialist

=> Word segmentation problem

Vietnamese word morphology

- Morphologic invariant
  - Some exceptions
    - Usage of some special characters in some case :

    Ex. “Bác sĩ” and “Bác sỹ” are the same meaning “Doctor”

    - Position of the tons

    Ex. “Hòa bình” or “hoà bình” are acceptable !
  - Prefix, suffix : “sự” , ‘hóa” : used infrequently

=> Word normalization is simpler

Vietnamese Word Category (POS : Part Of Speech)

- Dependent on context (can not recorgnize base on the word form like European Languages)

  1. "**thành công** (success) của dự án đã tạo tiếng vang lớn" (The success of the project created a big echo)

  2. "Anh ta đã **thành công** (succeed) trong nghiên cứu khoa học" (He have succeed in scientist research)

  3. "Buổi biểu diễn đã **thành công** (successful)" (The show was successful)

Vietnamese Text Retrieval

- What is better index terms for Vietnamese text ?
  - Linguistic unit “tiếng” : reuse of tokenization methods for European Languge (use white bank)
  - Word : need of word segmentation method
  - Noun phrase, concept : need of Vietnamese NLP tools as : Vietnamese POS tagger, Vietnamese Chunker
- Now : at the first steps
- How to evaluate Vietnamese IR ? Vietnamese test collection ?

Test collection

Document collections

- Monolingual Vietnamese Text Collection
  - New paper
  - Num of documents : 14.000
  - Size : 30Mb
  - Encoding : UTF-8
  - Format : TREC

Vietnamese Text Document sample

<TOP>
<NUM> 10</NUM>
<TITLE>
Thương mại Việt Mỹ
</TITLE>
<DESCRIPTION>
Các chính sách và hoạt động liên quan đến thương mại giữa Việt nam và Mỹ
</DESCRIPTION>
<NARRATIVE>
Các chính sách mới trong quan hệ thương mại hai nước, các cuộc tiếp xúc của các tổ chức thương mại của hai bên, các báo cáo về kết quả của sự hợp tác thương mại giữa hai nước. Các bài báo nói về các vấn đề trên được cho là liên quan.
</NARRATIVE>
</TOP>

<TOP>
<NUM> 10</NUM>
<TITLE>
Vietnam America Trading
</TITLE>
<DESCRIPTION>
The policies and activities relates to trading of Vietnam and America
<NARRATIVE>
The new policies in trading of two countries, the events are organized of trading organizations of two contries, the reports of trading cooperation Vietnam – America, the documents relate the subjects above are judged relevance.
</NARRATIVE>
</TOP>

Bilingual English-Vietnamese text collection

- Automatic mining from web
- Number of pair documents : 1468
- Size : 20Mb

Collection	N. of pair documents	Size
Vietnamese Law	336	15Mb
VOA (Voice of America)	1074	4Mb
US. Embassy	58	1Mb
	1468	20Mb

Sample

- **[P1]** ISRAELI TROOPS KILL 5 MORE PALESTINIANS IN GAZA
- **[P2]** AN ISRAELI HELICOPTER STRIKE HAS KILLED TWO PALESTINIAN TEENAGERS IN THE NORTHERN GAZA STRIP, AS THE MILITARY CONTINUES A MAJOR OFFENSIVE TO TRY TO STOP MILITANTS FROM FIRING ROCKETS INTO NEARBY JEWISH SETTLEMENTS
- **[P3]** RESIDENTS OF THE JABALYA REFUGEE CAMP SAY ONE OF THE TEENS WAS A MILITANT.
- **[P4]** ISRAEL'S MILITARY SAYS IT FIRED ON A GROUP OF GUNMEN WHO WERE TRYING TO PLANT A BOMB
- **[P5]** MEANWHILE, A PALESTINIAN BOY DIED FRIDAY FROM INJURIES SUSTAINED WHEN AN ISRAELI TANK FIRED ON THE REFUGEE CAMP LAST WEEK. A 10-YEAR-OLD GIRL WAS KILLED BY ISRAELI GUN FIRE IN THE SAME AREA TODAY
- **[P6]** IN A SEPARATE INCIDENT, OFFICIALS SAY PALESTINIAN MILITANTS SHOT AND KILLED A PALESTINIAN WORKING ON A FARM IN A JEWISH SETTLEMENT IN SOUTHERN GAZA
- **[P7]** MORE THAN 80 PALESTINIANS AND THREE ISRAELIS HAVE BEEN KILLED SINCE THE GAZA OFFENSIVE BEGAN LAST WEEK

- **[P1]** MÁY BAY TRỰC THĂNG ISRAEL BẮN CHẾT 2 THIẾU NIÊN PALESTINE TẠI DẢI GAZA
- **[P2]** MỘT MÁY BAY TRỰC THĂNG CỦA ISRAEL ĐÃ BẮN CHẾT 2 THIẾU NIÊN PALESTINE TẠI MIỀN BẮC DẢI GAZA KHI QUÂN ĐỘI TIẾP TỤC CUỘC HÀNH QUÂN LỚN ĐỂ NGĂN CHẶN CÁC PHẦN TỬ TRANH ĐẤU BẮN ROCKET VÀO CÁC KHU ĐỊNH CƯ DO THÁI
- **[P3]** CƯ DÂN TẠI TRẠI TỊ NẠN JABALYA NÓI RẰNG MỘT TRONG 2 THIẾU NIÊN VỪA KỂ LÀ MỘT PHẦN TỬ TRANH ĐẤU.
- **[P4]** QUÂN ĐỘI ISRAEL NÓI RẰNG HỌ BẮN VÀO MỘT NHÓM PHẦN TỬ VÕ TRANG ĐANG TÌM CÁCH GÀI BOM
- **[P5]** TRONG KHI ĐÓ MỘT BÉ TRAI PALESTINE TỪ TRẬN NGÀY HÔM NAY VÌ VẾT THƯƠNG DO MỘT XE TĂNG ISRAEL BẮN VÀO TRẠI TỊ NẠN HỒI TUẦN TRƯỚC
- **[P6]** HÔM NAY, MỘT BÉ GÁI BỊ THIỆT MẠNG NGÀY VÌ TRÚNG ĐẠN CỦA ISRAEL TRONG CÙNG KHU VỰC NÀY
- **[P7]** TRONG MỘT DIỄN BIẾN KHÁC, CÁC GIỚI CHỨC NÓI RẰNG CÁC PHẦN TỬ TRANH ĐẤU PALESTINE ĐÃ BẮN CHẾT MỘT NGƯỜI PALESTINE LÀM VIỆC TẠI MỘT CÔNG TRẠI TRONG MỘT KHU ĐỊNH CƯ DO THÁI TẠI MIỀN NAM DẢI GAZA
- **[P8]** HƠN 80 NGƯỜI PALESTINE VÀ 3 NGƯỜI ISRAEL ĐÃ THIỆT MẠNG KỂ TỪ KHI CUỘC HÀNH QUÂN CỦA ISRAEL BẮT ĐẦU HỒI TUẦN TRƯỚC

Search Topics

- 25 topics
- Choice from the themes of documents
- Criteria
  - Short topics
  - Long topics
  - Contain :
    - Simple word only
    - Simple word and compound word
    - Compound word only
- Format : TREC

Relevance Assessment

- Method : Pooling
- Used Systems :
  - SMART
  - Lemur
  - Terrier
- Pre-Works
  - We have modified these systems to work with Vietnamese character encoding UTF-8
  - Text collection pre-processing :
    - Vietnamese Word segmentation
    - connect the linguistic units of a word by _ (under score)
  - Modify tokenization module of Terrier

Relevance Assessment

- Use top 50 documents return by each system to make the pool

Experimentation

Experimentation purposes

- Test the different type of Vietnamese index term
- Test the indexing model for Vietnamese text

Experimentation scripts

- Test the types of Vietnamese index term
  - Linguistic unit : “uni-gram” : RUN_UNI
  - “Bi-gram” : two linguistic units adjunction : RUN_BI
  - Combination : uni-gram and lexicon : RUN_COM
- Test the indexing model (Use Lemur)
  - Okapi
  - Inquery
  - Language Model : KL-Divergence

Thanks you for your attention