Vietnamese Text Retrieval: Test Collection and First Experimentations

Ho Bao Quoc
Vietnam National University
HoChiMinh City University of Sciences
Where are we?
I am here !!!
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: ã, â, ê, ô,ù
• Usage six tons : (bằng), ‘ (sắc), ` (huyền), ?
  (hội) ~ (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: “surname”, ‘hóa” : used infrequently

=> Word normalization is simpler
Vietnamese Word Category
(POS : Part Of Speech)

• Dependent on context (can not recognize 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 Language (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
Vietnam America Trading
The policies and activities relates to trading of Vietnam and America
The new policies in trading of two countries, the events are organized of trading organizations of two countries, the reports of trading cooperation Vietnam – America, the documents relate the subjects above are judged relevance.
Bilingual English-Vietnamese text collection

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

<table>
<thead>
<tr>
<th>Collection</th>
<th>N. of pair documents</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vietnamese Law</td>
<td>336</td>
<td>15Mb</td>
</tr>
<tr>
<td>VOA (Voice of America)</td>
<td>1074</td>
<td>4Mb</td>
</tr>
<tr>
<td>US. Embassy</td>
<td>58</td>
<td>1Mb</td>
</tr>
<tr>
<td></td>
<td>1468</td>
<td>20Mb</td>
</tr>
</tbody>
</table>
ISRAELI TROOPS KILL 5 MORE PALESTINIANS IN GAZA

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.

RESIDENTS OF THE JABALYA REFUGEE CAMP SAY ONE OF THE TEENS WAS A MILITANT.

ISRAEL’S MILITARY SAYS IT FIRED ON A GROUP OF GUARDIANS WHO WERE TRYING TO PROTECT A BOMB.

MEANWHILE, A PALESTINIAN BOY DIED FROM INJURIES SUSTAINED WHEN AN ISRAELI TANK FIRED AT THE REFUGEE CAMP LAST WEEK. A 10-YEAR-OLD GIRL WAS KILLED BY ISRAELI GUNFIRE EARLIER TODAY.

A SEPARATE SHOT KILLED A PALESTINIAN WORKER AT A FARM IN A JEWISH SETTLEMENT IN THE SOUTH TODAY.

MORE THAN 80 PALESTINIANS AND THREE ISRAELIS HAVE BEEN KILLED SINCE THE GAZA OFFENSIVE BEGAN LAST WEEK.
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