API xử lý ngôn ngữ tự nhiên

Đồng Thị Ngân – FTI

Hà Nội August 4th

Goals

- Giúp các bạn **làm quen** với các API về xử lý ngôn ngữ tự nhiên
- **Chỉ** tập trung vào trình bày các API cho tiếng Việt. Các API cho tiếng Anh tương tự.
- **Không có ý định:**
  - trình bày chuyên sâu về:
    - các bài toán xử lý ngôn ngữ tự nhiên
    - các mô hình học máy
    - cách thu thập xử lý dữ liệu
    - ...
  - hay miêu tả cách xây dựng hệ thống

Overview

- Sentence segmentation
- Word segmentation
- Part-of-Speech tagging
- Named-entity recognition
- True-case
- Input-type
- Answer Type

Sentence Segmentation

- Là quá trình chia tách một đoạn văn bản thành các câu.
  - Input: một đoạn văn bản
  - Output: tập các câu trong đoạn văn bản đó, phân tách bởi ký tự xuống dòng
- Ví dụ:
  - Input: Mất cân đối đã thành sự thật. Mất trí nhớ, mai một ký ức là một nguy cơ nhãn tiền.
  - Output: 2 câu:
    - Mất cân đối đã thành sự thật.
    - Mất trí nhớ, mai một ký ức là một nguy cơ nhãn tiền.

Sentence Segmentation (cont.)

- Là *tiền đề* cho các bài toán xử lý cấp cao hơn như:
  - Tách từ (word segmentation)
  - Gán nhãn từ loại (POS tagging)
  - Phân tích cú pháp (Parser)
  - Dóng câu (sentence alignment)
  - Dịch tự động (auto-translation)
  - ...
- API:
  - http://54.255.200.131/fti-qa/nlp/en/segment/word?text=Mất cân đối đã thành sự thật.
  - Output:
    - {"status":"success","word_segmented_text":"Mất cân đối đã thành sự thật .**\n**"}

  → *mỗi câu được phân tách bởi ký tự "\n"*

Word Segmentation

- Là quá trình xác định ranh giới các từ trong câu văn
  - Input: một câu tiếng Việt
  - Output: câu được tách từ với
    - Các từ được phân tách bởi cách trắng
    - Các tiếng trong từ được phân tách bởi dấu gạch chân
- Ví dụ:
  - Input: ông Trương Gia Bình là ai
  - Output: ông Trương_Gia_Bình là ai

Word Segmentation (cont.)

- Được coi là bước xử lý quan trọng
- Là tiền đề cho các bài toán:
  - Nhận dạng thực thể
  - Gán nhãn từ loại
  - Phân loại văn bản
  - ...
- API:
  - http://54.255.200.131/fti-qa/nlp/en/segment/word?text=ông Trương Gia Bình là ai
  - Output
    - {"status":"success","word_segmented_text":"ông Trương_Gia_Bình là ai\n"}

Part-of-Speech Tagging

- Là việc xác định các chức năng ngữ pháp của từ trong câu.
- Ví dụ:
  - Input: Cô ấy cho tôi một quả cam.
  - Output: Cô/N ấy/P cho/E tôi/P một/M quả/Nc cam/N.
- Là bước cơ bản trước khi tiến hành
  - Phân tích sâu văn phạm
  - Các bước xử lý ngôn ngữ phức tạp hơn
    - Phân tích cú pháp, ...

Part-of-Speech Tagging (cont.)

- API:
  - http://54.255.200.131/fti-qa/nlp/vi/tagger?text=Cô ấy cho tôi một quả cam
  - Output:
    - {"status":"success","tagged_text":"Cô/N ấy/P cho/E tôi/P một/M quả/N cam/N"}

Part-of-Speech Tagging (cont.)

- Tập nhãn

Np - Proper noun	M - Numeral
Nc - Classifier	E - Preposition
Nu - Unit noun	C - Subordinating conjunction
N - Common noun	CC - Coordinating conjunction
V - Verb	I - Interjection
A - Adjective	T - Auxiliary, modal words
P - Pronoun	Y - Abbreviation
R - Adverb	Z - Bound morphemes
L - Determiner	X - Unknown

Named-entity recognition

- Là một nhiệm vụ con của trích xuất thông tin
  - tìm kiếm và phân loại các thành phần trong văn bản vào những loại xác định trước:
    - tên người, tổ chức, địa điểm, thời gian, số lượng, giá trị tiền tệ, phần trăm ...
- Ví dụ:
  - Input: tôi sống ở Hà Nội
  - Output: tôi sống ở <LOC>Hà Nội</LOC>

Named-entity recognition (cont.)

- Là bài toán đơn giản nhất trong số các bài toán trích chọn thông tin
- Là bước xử lý cơ bản nhất trước khi tính đến việc giải quyết các bài toán phức tạp hơn
- Tập Tags:
  - ***NUM***: number, ***DTIME***: date time, ***HUM***: human, ***LOC***: location, ***ORG***: organization
- API:
  - http://54.255.200.131/fti-qa/nlp/vi/ner?text=tôi sống ở Hà Nội
  - Output:
    - {"status":"success","name_entity_recognize_text":"tôi sống ở <LOC>Hà Nội</LOC>"}

True-case

- Sửa các lỗi viết hoa trong câu chữ thường
- Áp dụng với văn nói, khi các câu là đầu ra của một bộ nhận dạng tiếng nói
- Ví dụ:
  - Input: tôi sống ở hà nội
  - Output: Tôi sống ở Hà_Nội
- Giúp làm tăng độ chính xác của các hệ thống hiện thời:
  - gán nhãn từ loại,
  - nhận dạng tên,
  - …

  với *input là các câu chữ thường*

True-case (cont.)

- API:
  - http://54.255.200.131/fti-qa/nlp/vi/true-case?text=tôi sống ở hà nội
  - Output:
    - {"status":"success","truecased_text":"Tôi sống ở Hà_Nội"}

Input-type

- Thực hiện phân loại các câu đầu vào thành các kiểu câu định sẵn.
- Tập nhãn sử dụng:
  - Interrogative (**INT**): Câu hỏi (hỏi thông tin nào đó từ hệ thống)
  - Declarative (**DEC**): Câu trả lời hoặc khai báo (cung cấp thêm thông tin cho hệ thống)
  - Affirmative (**AFF**): Câu khẳng định (khẳng định, xác nhận điều gì, thông tin gì đó)
  - Imperative (**IMP**): Câu mệnh lệnh (muốn hệ thống thực hiện điều gì đó)
  - Commentary (**COM**): Câu nhận xét, bình luận (đưa ra nhận xét, bình luận nào đó)
  - Desire (**DES**): Câu thể hiện mong muốn (mong muốn, ý muốn thực hiện)
  - Advisory (**ADV**): Câu khuyên bảo (những câu mang tính khuyên bảo hệ thống)
  - Exclamatory (**EXC**): Câu cảm thán (miêu tả tình trạng, tâm trạng, cảm xúc)

Input-type (cont.)

- Ví dụ:
  - ai là chủ tịch fpt, INT
  - hôm nay thời tiết đẹp quá, EXC
  - hôm nay thời tiết tốt, DEC
  - không phải đâu, AFF
  - thèm ăn phở quá, DES
- API:
  - API: http://54.255.200.131/fti-qa/nlp/vi/sent-categorizer?text=thèm ăn phở quá
  - Output:
    - {"status":"success","senttype_text":"EXC"}

  Chú ý: Câu input đầu vào <u>không cần</u> có dấu câu

Answer Type

- Xác định loại câu trả lời cho một câu hỏi
- Tập tag:
  - **ORG**: organization, **HUM**: human, **LOC**: location, **DESC**: description, **THG**, **CONC**, **NUM**: number, **DTIME**: date time, **NET**: net, **EVT**: event
- Ví dụ:
  - Hiệu phó của trường ĐH FPT là những ai?, HUM
  - Tập đoàn FPT có những công ty con nào?, ORG
  - Tên đầy đủ hiện nay của FTG?, DESC
  - FPT Software/FSoft thành lập vào ngày tháng năm nào?, DTIME

Answer Type

- API:
  - http://54.255.200.131/fti-qa/nlp/vi/answer-type?question=hôm nay ngày bao nhiêu
  - Output:
    - {"status":"success","answer_type":"NUM"}

Conclusion

- Chúng tôi đã cung cấp 13 API về xử lý ngôn ngữ tự nhiên, bao gồm cả tiếng Anh và tiếng Việt
- Các hệ thống cung cấp:
  - nền tảng xử lý ngôn ngữ tự nhiên
  - Được tối ưu hóa cho bài toán tương tác người máy

Thank you!