Một số định hướng BTL

Lê Thanh Hương
Bộ môn Hệ thống Thông tin
Viện CNTT&TT

1

Một số định hướng BTL

- Tách từ (word segmentation)
- Gán nhãn từ loại (part-of-speech tagging)
- Phát hiện và sửa lỗi chính tả (spelling detection and correction)
- Phân tích cú pháp (syntactic parsing)
- Phân loại văn bản (text catergorization)
- Phân cụm văn bản (text clustering)

2

Một số định hướng BTL

- Trích rút thông tin (information extraction)
- Hệ thống hỏi đáp (Question answering)
- Chat bot
- Trợ lý ảo (visual assistant)
- Phân tích quan điểm (sentiment analysis)
- Nhận dạng thực thể (named entity recognition)
- Nhận dạng quan hệ giữa các thực thể (relation extraction)
- Trích rút sự kiện (event extection)
- Tóm tắt đơn/đa văn bản (text summarization/multi-document …)
- Gợi ý tin nhắn
- Hệ gợi ý (tin, sản phẩm,…) (recommender system)
- So sánh tin tức (text comparison/text alignment)
- Dịch máy (machine translation)
- Gióng hàng văn bản (1 ngôn ngữ, nhiều ngôn ngữ)

3

Phát hiện và sửa lỗi chính tả

- Ví dụ:
  - Lúi liền lúi, sông liền sông.
  - Con cáo và chùm ho
  - Giaỉ pháp, dichj máy
- Giải pháp:
  - Phân tích cấu tạo âm tiết
  - Từ điển
  - Ngram
  - Word2vec
  - RNN/CNN

4

Phân loại văn bản

- Học có giám sát: Naïve Bayes, SVM, …
- Các đặc trưng:
  - Ngram
  - POS
  - Từ điển
  - Từ viết tắt
  - Word2vec

5

Phân cụm văn bản

- Học không giám sát: kmeans, LSA/LDA…
- Các đặc trưng:
  - Ngram
  - POS
  - Từ viết tắt
  - Word2vec

6

Trích rút thông tin

- Nhận diện thực thể (named entity recognition – NER)
- Trích rút quan hệ giữa các thực thể (relation extraction)
- NER:
  - Gán nhãn sử dụng học máy: CRF
  - Các đặc trưng: ngram, POS, từ điển (tên riêng, từ viết tặt,…), định dạng từ (chữ viết hoa, là số, …)
- RE: phân loại

7

Hệ hỏi đáp

- Dữ liệu sẵn có: CSDL hoặc cơ sở tri thức thuộc lĩnh vực hỏi đáp
- 2 việc:
  - Phân tích câu hỏi
  - Tìm câu trả lời
- Phân tích câu hỏi:
  - Sử dụng template
  - Xây dựng văn phạm ngữ nghĩa để phân tích câu hỏi
  - Phân loại câu hỏi – intent detection (what/how many)/nhận diện thực thể - entity recognition (laptop-battery). Wit.ai là 1 tool theo hướng này
- Tìm câu trả lời: tìm trong dữ liệu có cấu trúc

8

Hệ hỏi đáp – tiếp

- Dữ liệu sẵn có: ngân hàng câu hỏi-đáp

Giải pháp:

- Tìm các câu hỏi trong ngân hàng câu hỏi tương đương câu đang hỏi
- Giải pháp:
  - So sánh độ tương đồng câu
  - So sánh/ tóm tắt các câu trả lời

9

Chatbot

- Hội thoại
- Tập các câu chào hỏi
- Phân loại câu hỏi – intent detection
- Nhận diện thực thể - entity recognition
- Sinh câu trả lời – tập luật

10

Phân tích quan điểm

- Mức:
  - văn bản, câu, dưới câu
  - Khía cạnh (aspect-based…)
- Thể loại: văn bản chính thống, mạng xã hội
- Cấp độ: (-1,1), (-1,0,1), (1-5)
- Giải pháp:
  - Khía cạnh: Named entity recognition
  - Học máy: Naïve Bayes, SVM,… RNN/CNN
  - Đặc trưng: ngram, POS, adj/adv, syntactic features, semantic features, sentiment features, negation features, …

11

Trích rút sự kiện

- Event(event,sbj,obj, time, place)
- Giải pháp:
  - Phân tích cú pháp, phân tích ngữ nghĩa
  - Nhận diện thực thể PER, TIME, LOC, ORG
  - Học máy

12

Tóm tắt đơn/đa văn bản

- Các đặc trưng:
  - Surface : đầu câu, đầu đoạn, độ dài câu
  - Nội dung: tf.idf, word2vec
  - Relevance: liên quan đến câu tiêu đề, PageRank
- Thuật toán:
  - Tính trọng số câu dựa trên các đặc trưng
  - Học máy dựa trên đặc trưng

13

Gợi ý tin nhắn

- Chuẩn hóa text
- Học dùng word modelling hoặc RNN/CNN

14

Hệ gợi ý tin

- Dựa trên sự tương đồng tin đang đọc – các tin khác

15

So sánh tin tức

- Các mức: so khớp/ tương tự ngữ nghĩa
- Giải pháp:
  - Đo độ tương đồng câu:
    - Các độ đo tương đồng cosin, leveistein dựa trên so khớp từ
    - Word2vec, doc2vec đo độ tương đồng dựa trên ngữ nghĩa
  - Định vị các đoạn tương đồng

16

Dịch máy

- Dựa trên luật chuyển đổi cây cú pháp giữa 2 ngôn ngữ
- Dựa trên gióng hàng câu, gióng hàng từ

17

Một số mã nguồn mở

- Stanford's Core NLP Suite (viết bằng Java): http://stanfordnlp.github.io/CoreNLP/
- Natural Language Toolkit (viết bằng Python): http://www.nltk.org/
- Apache Lucene and Solr: http://lucene.apache.org/
- Apache OpenNLP (viết bằng Java): http://opennlp.apache.org/
- Apache UIMA: https://uima.apache.org/
- GATE (General architecture for text engineering, viết bằng Java): https://gate.ac.uk/
- Lê Hồng Phương: http://mim.hus.vnu.edu.vn/phuonglh/softwares
- http://viet.jnlp.org/
- https://ongxuanhong.wordpress.com/2016/02/06/gioi-thieu-cac-cong-cu-xu-ly-ngon-ngu-tu-nhien/
- Underthesea: https://github.com/magizbox/underthesea

18

Một số hướng nghiên cứu

- Chatbot, question answering, visual assistant
- Phát hiện sao chép
- Tóm tắt văn bản
- Phát hiện sự kiện
- Trích rút thông tin
- Phân tích quan điểm
- Phân tích cú pháp

19