Trường Công nghệ Thông tin và Truyền thông
Khoa Công nghệ thông tin

XỬ LÝ NGÔN NGỮ TỰ NHIÊN

GVHD: Lâm Nhựt Khang
lnkhang@ctu.edu.vn

GIỚI THIỆU HỌC PHẦN

Nội dung học phần

1. Giới thiệu NLP
2. Mô hình ngôn ngữ
3. Vector ngữ nghĩa và nhúng từ
4. Gán nhãn từ loại
5. Phân tích cú pháp
6. Rút trích thông tin (option)
7. Case study (option)

Điểm thành phần

1. 02 project nhỏ: 15% x 2
2. 01 bài tập tại lớp: 10%
3. 01 project lớn: đề cương 10% + báo cáo cuối học phần 50%

Tài liệu tham khảo

Bài giảng được tổng hợp từ nhiều nguồn:

1. Dan Jurafsky and James H. Martin, 2024, Speech and Language Processing (3rd ed. draft), https://web.stanford.edu/~jurafsky/slp3/
2. Jacob Eisenstein, 2018, Natural Language Processing https://cseweb.ucsd.edu/~nnakashole/teaching/eisenstein-nov18.pdf
3. The Stanford Natural Language Processing Group https://nlp.stanford.edu/
4. Natural Language Toolkit https://www.nltk.org/

CHƯƠNG 1

GIỚI THIỆU TỔNG QUAN

Nội dung Chương

1. Giới thiệu tổng quan NLP
2. Các mức độ ngôn ngữ
3. Tài nguyên ngôn ngữ
4. Hướng tiếp cận
5. Phương pháp đánh giá trong NLP
6. Vì sao NLP khó
7. Công cụ tách từ tiếng Việt

7

Nội dung Chương

1. **Giới thiệu tổng quan NLP**
2. Các mức độ ngôn ngữ
3. Tài nguyên ngôn ngữ
4. Hướng tiếp cận
5. Phương pháp đánh giá trong NLP
6. Vì sao NLP khó
7. Công cụ tách từ tiếng Việt

8

Ngôn ngữ tự nhiên

Trong khoa học máy tính, ngôn ngữ mà con người sử dụng để giao tiếp với nhau được gọi là ngôn ngữ tự nhiên (*natural language*).

Udacity India. 2018. "What are the current hot topics in Natural Language Processing?" Medium, September 28. Accessed 2019-10-09.

Xử lý ngôn ngữ tự nhiên (NLP)

- NLP là lĩnh vực phân tích, thiết kế thuật toán với input và output là dữ liệu không có cấu trúc hoặc ngôn ngữ tự nhiên. [Goldberg, 2017]
- Xử lý ngữ tự nhiên tập trung phân tích thiết kế các thuật toán và cách thức biểu diễn để xử lý ngôn ngữ tự nhiên của con người. [Eisenstein, 2018]

Goldberg, Y., 2017. Neural network methods for natural language processing. Synthesis Lectures on Human Language Technologies, 10(1), pp.1-309.

Eisenstein, J. (2018). Natural language processing. Technical report, Georgia Tech

NLP và CL

- Ngôn ngữ học tính toán (Computational Linguistics - CL) nghiên cứu xử lý, tính toán trên nền tảng ngôn ngữ của con người:
  - Bằng cách nào chúng ta hiểu ngôn ngữ? (How do we understand language?)
  - Con người tạo ra ngôn ngữ như thế nào? (How do we produce language?)
  - Chúng ta học ngôn ngữ như thế nào? How do we learn language?
- NLP nghiên cứu, phát triển các phương pháp để giải quyết các vấn đề liên quan đến ngôn ngữ [Johnson, 2014].
  - Nhận dạng giọng nói (automatic speech recognition)
  - Dịch máy (machine translation)
  - Rút trích thông tin (information extraction from documents)

Vì sao máy tính gặp khó khăn với NLP ?

[Taylor, 2018]:

- Computers have mostly been dealing with structured data.
- This is data that's organized, indexed and referenced, often in databases

In NLP, we often deal with unstructured data [Morikawa, 2018]:

- Human languages are quite unlike the precise and unambiguous nature of computer languages.
- Human languages have plenty of complexities such as ambiguous phrases, colloquialisms, metaphors, puns, or sarcasms

Taylor, Christine. 2018. Structured vs. Unstructured Data. Datamation, March 28. Accessed 2019-06-09.

Morikawa, Rei. 2018. What is the difference between natural language processing (NLP) and natural language understanding (NLU). Quora, October 16. Accessed 2019-06-12.

Vì sao máy tính gặp khó khăn với NLP ?

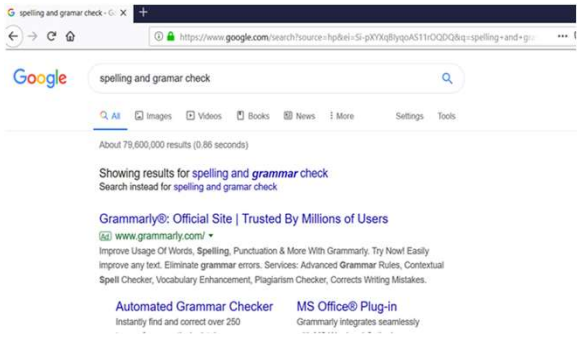
NLP phân tích văn bản không cấu trúc để rút trích thông tin [Waldron, 2015]

Waldron, Mike. 2015. "Structured vs Unstructured Data: Exploring an Untapped Data Reserve." AYLIEN, April 15. Accessed 2019-06-09.

LNK_13

Ứng dụng NLP

Spell and Grammar Checking
- Kiểm tra chính tả và ngữ pháp
- Gợi ý sửa lỗi



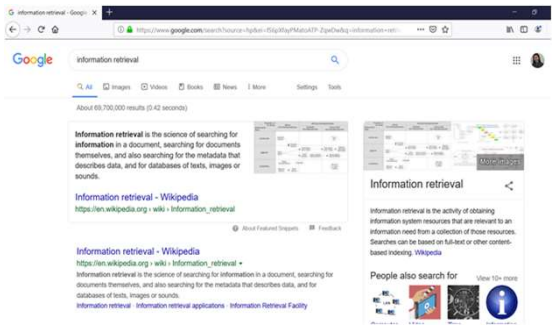
Word prediction
- Dự đoán từ tiếp theo mà người dùng có khả năng nhập vào

LNK_14

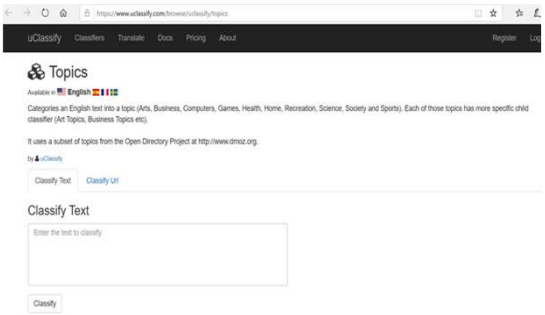
Ứng dụng NLP

Information Retrieval

- Tìm kiếm những thông tin có nội dung phù hợp với câu hỏi của người dùng



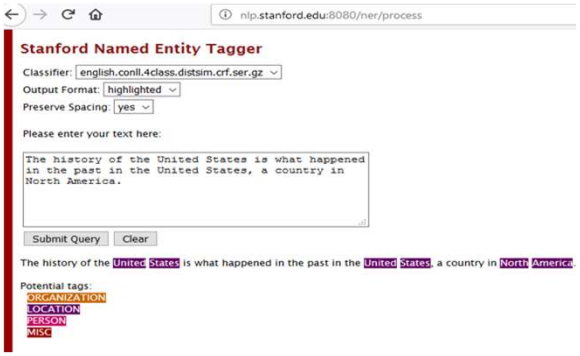
Text Categorization



Ứng dụng NLP

Summarization

Ứng dụng NLP



Information Extraction

- Extracting important concepts from texts and assigning them to slot in a certain template
- Includes named-entity recognition
- [Venali Sonone, 2018] Information Extraction refers to the automatic extraction of structured information such as entities, relationships between entities, and attributes describing entities from unstructured sources.

Venali Sonone. 2018. "Tutorial Series on NLP: Information Extraction tasks". August 29. Accessed 2019-10-19. https://medium.com/@venali/tutorial-series-on-nlp-information-extraction-tasks-99cd8309e2ef

Ứng dụng NLP

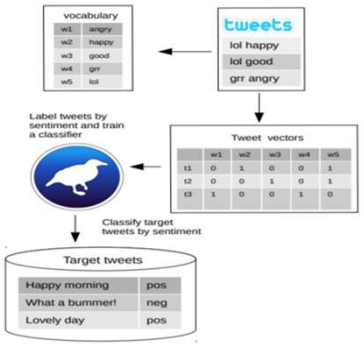
Sentiment Analysis

Main Problem: Message-level Polarity Classification (MPC)

- Automatically classify a sentence to classes positive, negative, or neutral.

Mohammad, S. M., Kiritchenko, S., and Zhu, X. (2013). Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013).

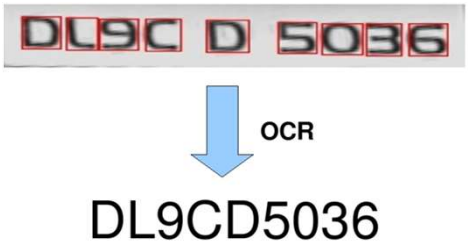
Sentiment Classification via Supervised Learning and BoWs Vectors



Felipe Bravo-Marquez, 2019, Natural Language Processing Introduction (slide)

Ứng dụng NLP

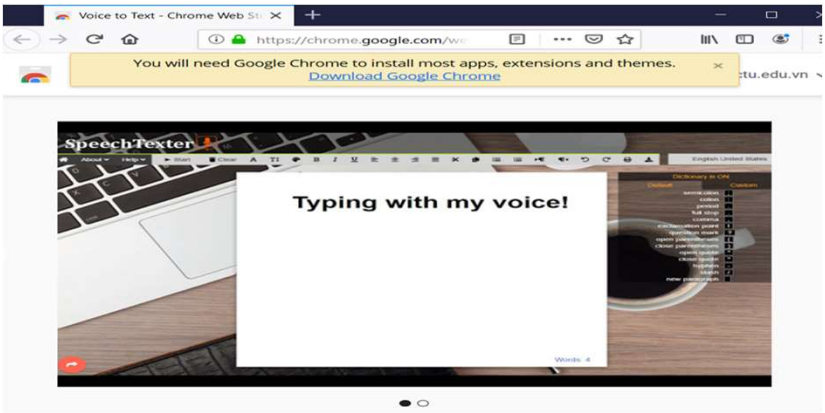
Optical Character Recognition



LNK_19

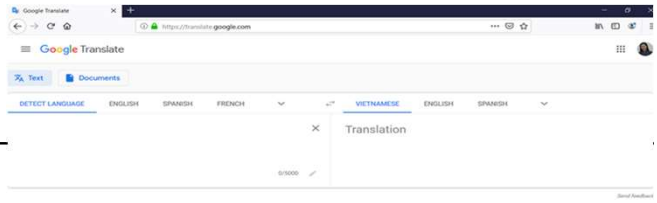
Ứng dụng NLP

Speech recognition



LNK_20

Ứng dụng NLP



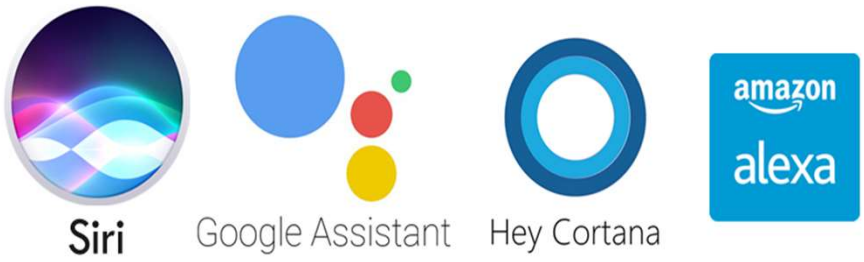
Machine Translation

- Facebook uses machine translation to automatically translate posts and comments.
- Google Translate processes 100 billion words a day.
- To connect sellers and buyers across language barriers, eBay is using machine translation

Le, James. 2018. The 7 NLP Techniques That Will Change How You Communicate in the Future (Part I). Heartbeat, via Medium, June 06. Accessed 2019-10-09.

Ứng dụng NLP

Question Answering



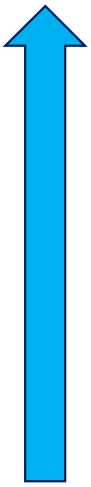
Ứng dụng NLP

Độ "khó"

Cấp độ khó (vẫn còn "khó")
•Hỏi đáp (Question answering)
•Tóm tắt (Summarization)
•Hệ thống hội thoại (Dialog systems)

Cấp độ trung bình (đang tiến triển tốt)
•Tìm kiếm thông tin (Information retrieval)
•Phân tích cảm xúc (Sentiment analysis)
•Dịch máy (Machine translation)
•Rút trích thông tin (Information extraction)

Cấp độ dễ (hầu như đã được giải quyết)
•Kiểm tra chính tả và ngữ pháp
•Một vài tác vụ phân loại text (text categorization tasks)
•Một vài tác vụ liên quan NER (named-entity recognition tasks)



Nội dung Chương

1. Giới thiệu tổng quan NLP
2. **Các mức độ ngôn ngữ**
3. Tài nguyên ngôn ngữ
4. Hướng tiếp cận
5. Phương pháp đánh giá trong NLP
6. Vì sao NLP khó
7. Công cụ tách từ tiếng Việt

Các mức độ ngôn ngữ

Ngôn ngữ học nghiên cứu các mức độ hoặc khía cạnh khác nhau của cấu trúc ngôn ngữ, cũng như cách thức mà cấu trúc ngôn ngữ tương tác với nhận thức của con người và xã hội [Bender, 2013].

- Ngữ âm học (phonetics): the sounds of human language.
- Âm vị học (phonology): sound systems in human languages.
- Hình thái học (morphology): the formation and internal structure of words.
- Cú pháp (syntax): The study of the formation and internal structure of sentences.
- Ngữ nghĩa (semantics): the study of the meaning of sentences
- Ngữ dụng (pragmatics): the study of the way sentences with their semantic meanings are used for particular communicative goals.

Bender, E.M., 2013. Linguistic fundamentals for natural language processing: 100 essentials from morphology and syntax. Synthesis lectures on human language technologies, 6(3), pp.1-184.

Ngữ âm học & Âm vị học

Johnson, M. (2014). Introduction to computational linguistics and natural language processing (slides). Machine Learning Summer School

Ngữ âm học (phonetics): nghiên cứu về âm thanh của một ngôn ngữ [Johnson, 2015]. Hệ thống ngữ âm học quốc tế là IPA (International Phonetic Alphabet): alphabetic system of phonetic notation.

Âm vị học (phonology): âm vị là đơn vị ngữ âm nhỏ nhất có chức năng khu biệt (Ân, 2017). Âm vị học nghiên cứu âm thanh của ngôn ngữ cụ thể. Âm vị học gồm có nguyên âm (*vowels*) và phụ âm (*consonants*). Tiếng Việt có 16 âm vị là nguyên âm và 23 âm vị là phụ âm (http://vnlp.net). Từ điển Việt - Việt của Hồ Ngọc Đức định nghĩa:

- Nguyên âm là âm phát từ những dao động của thanh quản, tự nó đứng riêng biệt hay phối hợp với phụ âm thành tiếng trong lời nói, phụ âm có thể ở trước hay ở sau hoặc cả trước lẫn sau: a, e, i, o... Ví dụ âm vị là nguyên âm: /i/ (in, ý chí, ….) hay /u/ (lu, tu hú, …)
- Phụ âm là âm phát từ thanh quản qua miệng, chỉ khi phối hợp với nguyên âm mới thành tiếng trong lời nói: b, c, d, l... Ví dụ âm vị là phụ âm: /f/ (phố, phường, …) hay /t/ (tin, tôi, …)

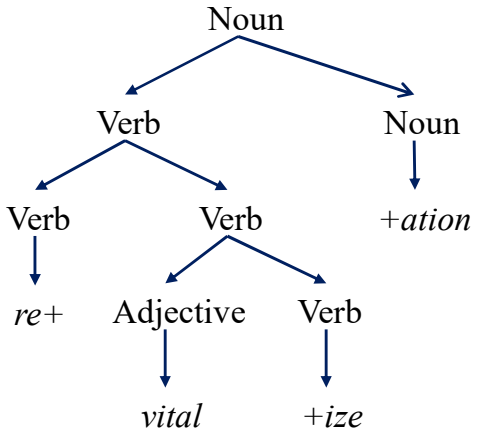
http://vnlp.net/ti%E1%BA%BFng-vi%E1%BB%87t-c%C6%A1-b%E1%BA%A3n/h%E1%BB%87-th%E1%BB%91ng-am-v%E1%BB%8B/

Hình thái học (morphology)

Hình thái học nghiên cứu về cấu trúc bên trong (internal structure) của từ

Ví dụ: re+structur+ing, un+remark+able, re+vital+ize+ation [Johnson, 2014]

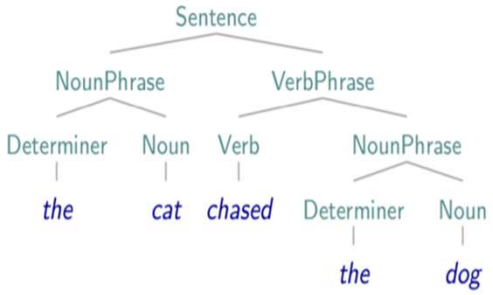
Theo [Phạm Thị Tươi, 2012] thì tri thức hình thái học là tri thức về "các từ được cấu trúc như thế nào từ những đơn vị nghĩa cơ bản hơn, được gọi là hình vị (morpheme)"



Cú pháp (syntax)

- Cú pháp nghiên cứu về mối quan hệ cấu trúc giữa các từ trong câu
- Cú pháp nghiên cứu cách các từ kết nối lại (combine) để hình thành cụm từ (phrases) và câu (sentences) [Johnson, 2014]
- Phân tích cú pháp để có thể xác định "**who did what to whom**", giúp hiểu một câu (understand a sentence).

Johnson, M. (2014). Introduction to computational linguistics and natural language processing (slides). Machine Learning Summer School



LNK_29

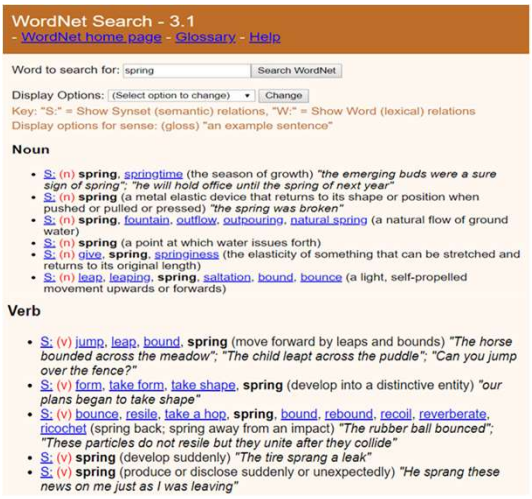
Ngữ nghĩa (semantics)

Nghiên cứu nghĩa của từ, và cách kết hợp (combine) các từ để tạo nghĩa trong câu

Quan hệ từ vựng (lexical relations): mối quan hệ giữa các từ [Yule, 2016].

- Synonymy: fall & autumn
- Hypernymy & hyponymy (is a): animal & dog
- Meronymy (part of): finger & hand
- Homonymy: fall (verb & reason)
- Antonymy: big & small

Yule, G. (2016). The study of language. Cambridge university press



LNK_30

Ngữ dụng (pragmatics)

Ngữ dụng (pragmatics) nghiên cứu cách thức ngữ cảnh (context) ảnh hưởng đến nghĩa (meaning) trong từng trường hợp cụ thể [Fromkin et al., 2018].

Fromkin, V., Rodman, R., and Hyams, N. (2018). An introduction to language. Cengage Learning

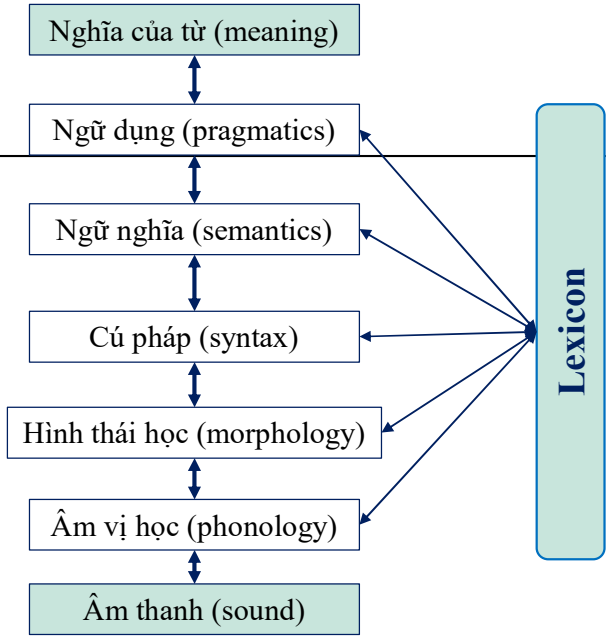
Ví dụ: Xét câu "Ông già đi nhiều rồi":

- Trong ngữ cảnh những người bạn thân lâu ngày gặp lại nhau thì có thể hiểu là một người đã bị già (lão hóa) nhiều rồi hay câu có thể được phân tách thành "Ông / già đi/ nhiều rồi".
- Trong ngữ cảnh nói về một cụ ông đã đi bộ nhiều thì câu có thể được phân tách thành "Ông già / đi/ nhiều rồi"

Lexicon

Mỗi ngôn ngữ có một *lexicon* gồm danh sách (list) các hình vị (morphemes) và từ (words) cung cấp các thông tin như cách phát âm, hình thái học, cú pháp, ngữ nghĩa,…

Johnson, M. (2014). Introduction to computational linguistics and natural language processing (slides). Machine Learning Summer School



Nội dung Chương

1. Giới thiệu tổng quan NLP
2. Các mức độ ngôn ngữ
3. **Tài nguyên ngôn ngữ**
4. Hướng tiếp cận
5. Phương pháp đánh giá trong NLP
6. Vì sao NLP khó
7. Công cụ tách từ tiếng Việt

Tài nguyên ngôn ngữ

Từ điển:

- Landau [11] định nghĩa từ điển "là một danh sách các các entry được sắp xếp theo các đơn vị từ vựng (lexical unit).
- Mỗi entry thường bao gồm một đơn vị từ vựng, giải nghĩa của từ, từ loại, phát âm, ví dụ minh họa cách sử dụng từ và một số thông tin khác.
- Đơn vị từ vựng thường là một từ (single word) trong khi các giải nghĩa của từ thường là một cụm từ (phrase).
- Ví dụ từ điển:
  - Từ điển đơn ngữ (từ điển tiếng Anh Oxford),
  - Từ điển song ngữ (từ điển Anh-Việt của Hồ Ngọc Đức),
  - Từ điển đồng nghĩa (từ điển các từ đồng nghĩa của Merriam-Webster [12]), hoặc
  - Các từ điển chuyên ngành (từ điển chuyên ngành Luật Black's Law Dictionary [13]).

Tài nguyên ngôn ngữ

Thesaurus:

- Kilgarriff [14] mô tả theasaurus khá giống từ điển đồng nghĩa với các từ có ngữ nghĩa tương tự nhau;
- Điểm khác biệt giữa thesaurus và từ điển đồng nghĩa là các khái niệm trong thesaurus có mối liên hệ lẫn nhau (*relationship*).
- Ví dụ: Roget's International Thesaurus và Open Thesaurus

Tài nguyên ngôn ngữ

Corpus là kho ngữ liệu văn bản (text) có cấu trúc.

- Brown corpus (dữ liệu tiếng Anh)
  - *https://www.kaggle.com/nltkdata/brown-corpus*
- Reuters corpus (dữ liệu ở hơn 300 ngôn ngữ) đều đã được tích hợp bên trong NLTK
  - *https://trec.nist.gov/data/reuters/reuters.html*
- British National Corpus (còn được gọi là BNC corpus)
  - *http://www.natcorp.ox.ac.uk/*
- Corpus of Contemporary American English (CoCa corpus)
  - *https://www.english-corpora.org/coca/*
- WordBanks Online
  - *https://wordbanks.harpercollins.co.uk/Docs/WBO/WordBanksOnline_English.html*
- IntelliText
  - *http://corpus.leeds.ac.uk/it/*

Tài nguyên ngôn ngữ

Penn TreeBank:

- Là một kho dữ liệu văn bản lớn với 4 định dạng: dữ liệu thô, dữ liệu đã được gán nhãn từ loại POS, dữ liệu đã được phân tích cú pháp và dữ liệu tổng hợp chứa cả POS và đã được phân tích cú pháp.
- https://catalog.ldc.upenn.edu/LDC99T42

Tài nguyên ngôn ngữ

WordNet:

- Là một cơ sở dữ liệu từ vựng, trong đó các từ có cùng ngữ nghĩa (*cognitive synonym*) được nhóm lại thành các nhóm từ đồng nghĩa.
- Mỗi nhóm từ đồng nghĩa được gọi là một *synset*.
- Các từ trong cùng một synset có ngữ nghĩa giống nhau trong một ngữ cảnh nào đó. Một từ đa nghĩa sẽ thuộc nhiều hơn một synset.
- WordNet lớn nhất trên thế giới là Princeton Wordnet, là Wordnet tiếng Anh, được xây dựng thủ công bởi các chuyên gia từ những năm 1995.
- PWN được xem là cơ sở dữ liệu từ vựng lớn nhất thế giới, vẫn đang trong quá trình xây dựng và hoàn thiện. WordNet cũng đã được tích hợp vào NLTK.

https://wordnet.princeton.edu/

Tài nguyên ngôn ngữ

VerbNet:

- Là một mạng động từ tiếng Anh, được ánh xạ tới các tài nguyên từ vựng khác như WordNet, PropBank và FrameNet.
- VerbNet được tổ chức thành các lớp động từ (*verb classes*) mở rộng từ các lớp Levin thông qua sàng lọc và bổ sung các lớp con để có thể đạt được sự kết hợp cú pháp và ngữ nghĩa giữa các thành viên của một lớp.
- Mỗi lớp động từ trong VerbNet gồm có một tập mô tả cú pháp và một tập mô tả ngữ nghĩa của động từ.

https://wordnet.princeton.edu/

Nội dung Chương

1. Giới thiệu tổng quan NLP
2. Các mức độ ngôn ngữ
3. Tài nguyên ngôn ngữ
4. **Hướng tiếp cận**
5. Phương pháp đánh giá trong NLP
6. Vì sao NLP khó
7. Công cụ tách từ tiếng Việt

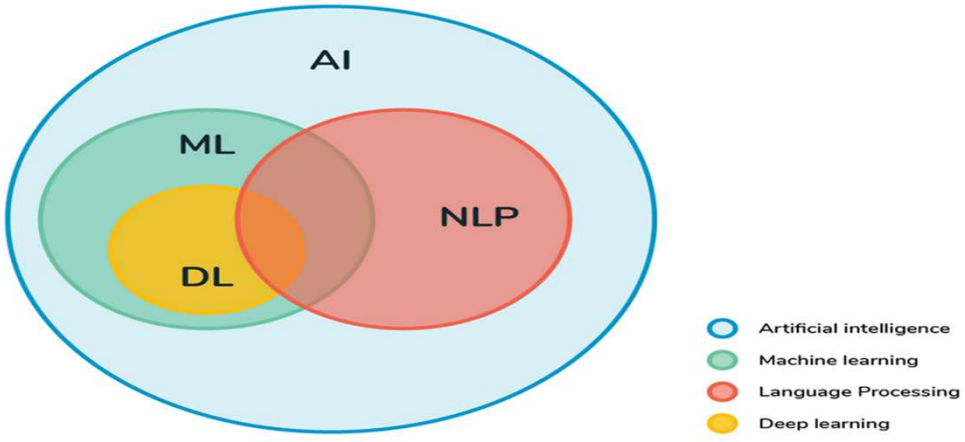
AI, ML, DL, và NLP

Theo Vajjala và các cộng sự [18]:

- Trí tuệ nhân tạo (*Artificial Intelligence - AI*) là một nhánh của khoa học máy tính với mục tiêu là xây dựng các hệ thống thông minh có khả năng thực hiện các tác vụ đòi hỏi trí tuệ của con người;
- Máy học (*Machine Learning - ML*) là một phần của AI, tập trung vào phát triển các thuật toán có khả năng "học" từ dữ liệu để tự thực hiện các tác vụ mà không cần các quy tắc cụ thể;
- Học sâu **(***Deep Learning - DL***)** là một nhánh của ML, dựa trên các mạng nơron nhân tạo để cải thiện trí thông minh, giúp giải quyết tác vụ tốt hơn;
- NLP cũng là một nhánh con của AI, sử dụng dụng các kỹ thuật ML và DL để hiểu, phân tích, và tạo ra ngôn ngữ tự nhiên.

Tóm lại, ML, DL, và NLP đều là các nhánh nhỏ của AI và có mối liên hệ với nhau.

AI, ML, DL, và NLP



NLP, NLU, và NLG

[Harper, 2018]:

- NLU is about analysis.
  - Example: Sentiment analysis and semantic search are examples of NLU
- NLG is about synthesis.
  - Example: Captioning an image or video is mainly an NLG task since input is not textual.
- An NLP application may involve one or both.
  - Example: Text summarization and chatbot are applications that involve NLU and NLG.

Harper, Jelani. 2018. "2019 Trends in Natural Language Processing." AI Business, October 16. Updated 2019-01-09. Accessed 2019-10-19.

LNK_43

NLP, NLU, và NLG

[Nuseibeh, 2018] Natural Language Understanding (NLU):

- This involves converting speech or text into useful representations on which analysis can be performed.
- The goal is to resolve ambiguities, obtain context and understand the meaning of what's being said. S
- NLP is about text parsing and syntactic processing while NLU is about semantic relationships and meaning

[Morikawa, 2018] NLU tackles the complexities of language beyond the basic sentence structure

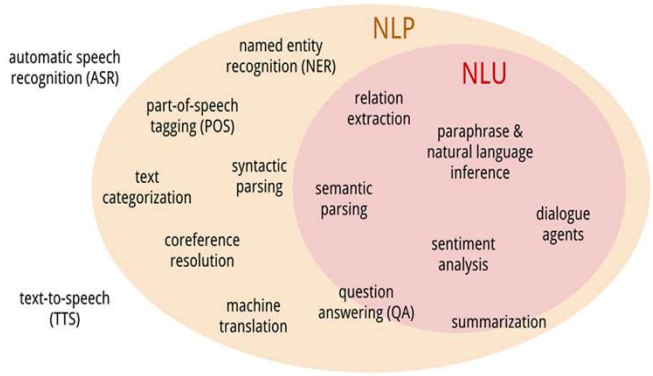
Natural Language Generation (NLG): *https://devopedia.org/natural-language-processing#Le-2018*

- Given an internal representation, this involves selecting the right words, forming phrases and sentences. Sentences need to ordered so that information is conveyed correctly

Nuseibeh, Rajai. 2018. "NLP; NLU and NLG Conversational Process Automation Chatbots explained." botique.ai, via Medium, November 27. Accessed 2019-10-19
Morikawa, Rei. 2018. "What is the difference between natural language processing (NLP) and natural language understanding (NLU)?" Quora, October 16. Accessed 2019-10-19

LNK_44

NLP, NLU, và NLG



DataFlair. 2018. "What is Natural Language Processing in Artificial Intelligence?" DataFlair, January 24. Accessed 2019-10-19.

Hướng tiếp cận NLP

- Các tác vụ NLP có thể được giải quyết bằng nhiều phương pháp khác nhau,
- Tuy nhiên chúng ta có thể nhóm các phương pháp này vào 3 hướng tiếp cận chính:
  - Heuristic,
  - Học máy, và
  - Học sâu.

Hướng tiếp cận NLP

Hướng tiếp cận Heuristic:

- Các hệ thống NLP cũng sử dụng các tri thức của chuyên gia để giải quyết tác vụ cụ thể. Các nguồn tài nguyên từ vựng ở cấp độ từ (*word-level*) như từ điển, thesaurus, và WordNet thường được sử dụng để giải quyết các bài toán NLP.
- Ngoài ra, các hệ thống dựa trên tập luật (*rule-based*) có thể chỉ sử dụng tài nguyên cấp độ từ hoặc kết hợp với các dạng thông tin khác như biểu thức chính quy (*Regular expressions*) và văn phạm phi ngữ cảnh.

Hướng tiếp cận NLP

Hướng tiếp cận học máy:

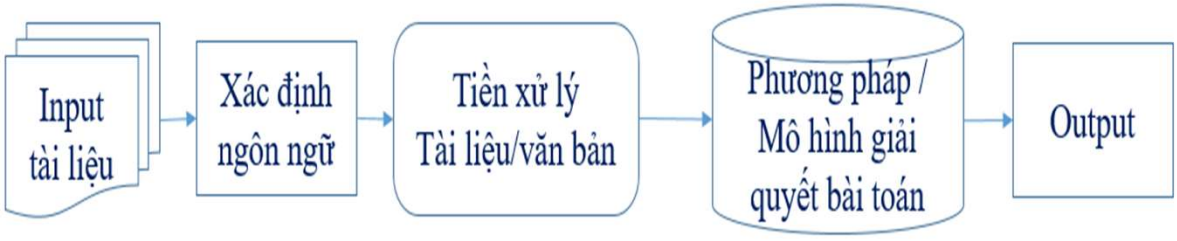
- Tương tự như các loại dữ liệu khác như hình ảnh và âm thanh, các kỹ thuật học máy cũng được sử dụng cho dữ liệu văn bản.
- Các kỹ thuật học máy có giám sát như phân lớp và hồi quy (*regression*) được sử dụng phổ biến trong các tác vụ NLP, ví dụ như bài toán phân lớp văn bản hoặc dự đoán giá vàng.
- Các phương pháp học máy có giám sát hay không giám sát trong NLP đều gồm 3 bước cơ bản là:
  - Rút trích đặc trưng từ văn bản,
  - Sử dụng các biểu diễn đặc trưng để huấn luyện mô hình, và
  - Đánh giá và cải tiến mô hình.
- Một số mô hình máy học có giám sát được sử dụng trong NLP có thể kể đến như Naïve Bayes, máy học vector hỗ trợ SVM, mô hình Markov ẩn HMM, và Conditional random fields.

Hướng tiếp cận NLP

Hướng tiếp cận học sâu:

- Một số mô hình học sâu được sử dụng trong NLP có thể kể đến như Recurrent neural networks (RNN), Long short-term memory (LSTM), Convolutional neural networks (CNN), Transformer, và AutoEncoder.
- Các mô hình học sâu tồn tại một số nhược điểm:
  - Khả năng bị overfitting khi sử dụng tập dữ liệu có kích thước hạn chế,
  - Học với ít dữ liệu và tạo dữ liệu giả (synthetic data generation) chưa được áp dụng thành công trong NLP,
  - Gặp khó khăn trong tương thích miền dữ liệu (ví dụ: mô hình được huấn luyện để tóm tắt các bài báo online, thì khi áp dụng vào tóm tắt các bài báo khoa học không đạt hiệu quả cao),
  - Khả năng kiểm soát và giải thích được các mô hình DL trong tác vụ NLP cũng còn hạn chế,
  - Phí tổn về tiền bạc và thời gian để xây dựng các giải pháp cho tác vụ NLP sử dụng DL là lớn.

Các bước cơ bản để giải quyết



Sentence splitting

VNEXPRESS
Báo tiếng Việt nhiều người xem nhất
Quảng cáo

Video | Thời sự | Góc nhìn | Thế giới | Kinh doanh | Giải trí | Thể thao | Pháp luật | Giác
Doanh nghiệp | Bất động sản | Ebank | Thương mại điện tử | Hàng hóa | Tiền c
Kinh doanh | Quốc tế

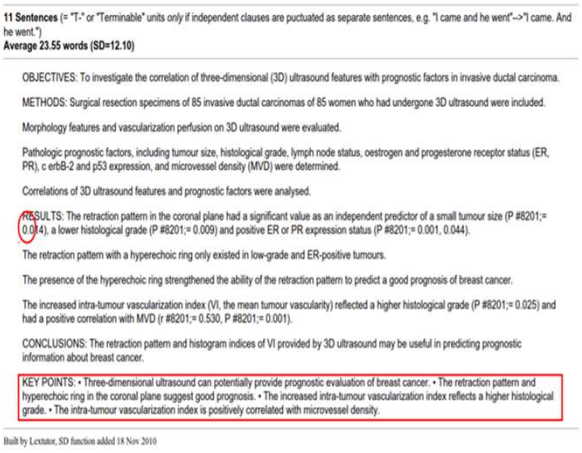
Thứ sáu, 18/10/2019, 01:05 (GMT+7)

Thu nhập bao nhiêu để vào nhóm 1% giàu nhất Mỹ?

Nhóm 1% người giàu nhất Mỹ kiếm được hơn 515.000 USD một năm và đóng thuế thu nhập cao hơn 90% dân số có thu nhập top dưới.

Một trong các định nghĩa về sự giàu có là nằm trong nhóm 1% người giàu nhất nước. Với người dân Mỹ, mục tiêu này đang ngày càng xa vời. Theo số liệu được Sở thuế Mỹ công bố tuần này, nhóm 1% người giàu nhất Mỹ có thu nhập hơn 515.000 USD năm 2017. Con số này tăng 7,2% so với năm trước đó. Kể từ năm 2011, ngưỡng này đã liên tục đi lên.

Trong khi đó, để gia nhập nhóm 0,1% người giàu nhất, bạn sẽ phải kiếm được 2,4 triệu USD năm 2017 - tăng 38% kể từ năm 2011. Còn nhóm cao nhất - 0,001%, với 1.433 đại diện năm 2017, mỗi người có thu nhập ít nhất 63,4 triệu USD.



11 Sentences (= "T-" or "Terminable" units only if independent clauses are puctuated as separate sentences, e.g. "I came and he went"-->"I came. And he went.")
Average 23.55 words (SD=12.10)

OBJECTIVES: To investigate the correlation of three-dimensional (3D) ultrasound features with prognostic factors in invasive ductal carcinoma.

METHODS: Surgical resection specimens of 85 invasive ductal carcinomas of 85 women who had undergone 3D ultrasound were included.

Morphology features and vascularization perfusion on 3D ultrasound were evaluated.

Pathologic prognostic factors, including tumour size, histological grade, lymph node status, oestrogen and progesterone receptor status (ER, PR), c erbB-2 and p53 expression, and microvessel density (MVD) were determined.

Correlations of 3D ultrasound features and prognostic factors were analysed.

RESULTS: The retraction pattern in the coronal plane had a significant value as an independent predictor of a small tumour size (P #8201;= 0.044), a lower histological grade (P #8201;= 0.009) and positive ER or PR expression status (P #8201;= 0.001, 0.044).

The retraction pattern with a hyperechoic ring only existed in low-grade and ER-positive tumours.

The presence of the hyperechoic ring strengthened the ability of the retraction pattern to predict a good prognosis of breast cancer.

The increased intra-tumour vascularization index (VI, the mean tumour vascularity) reflected a higher histological grade (P #8201;= 0.025) and had a positive correlation with MVD (r #8201;= 0.530, P #8201;= 0.001).

CONCLUSIONS: The retraction pattern and histogram indices of VI provided by 3D ultrasound may be useful in predicting prognostic information about breast cancer.

KEY POINTS: • Three-dimensional ultrasound can potentially provide prognostic evaluation of breast cancer. • The retraction pattern and hyperechoic ring in the coronal plane suggest good prognosis. • The increased intra-tumour vascularization index reflects a higher histological grade. • The intra-tumour vascularization index is positively correlated with microvessel density.

Built by Lextutor, SD function added 18 Nov 2010

Lemmatization và Stemming

Lemmatization:

- Thường thực hiện phân tích hình thái của từ và chuyển từ về dạng từ gốc (lemma) đúng chính tả hoặc từ có trong từ điển.
- Các từ có wordform khác nhau nhưng có cùng lemma.
- Các wordform “runs”, “running”, “ran” sau quá trình lemmatize sẽ chuyển về lemma “run”.
- NLTK toolkit cung cấp WordNet Lemmatizer dựa trên WordNet để thực hiện lemmatize từ.
- Phương pháp thực hiện lemmatize bao gồm phân tích hình thái học của từ (không nằm trong phạm vi của học phần này).

Lemmatization và Stemming

Stemming:

- Thường sử dụng thuật toán Heuristic để loại bỏ phần cuối của từ để đưa chúng về dạng "từ gốc" mà từ gốc đó có thể không phải là từ đúng chính tả.
- Các phương pháp stemming phổ biến gồm có Porter's stemmer [21], Lovins stemmer [22] và Paice/Husk stemmer [23].
- Stemming có khả năng sẽ ánh xạ một nhóm từ về cùng một "từ gốc",
- Ví dụ Porter stemmer sẽ chuyển các từ "trouble", "troubling", "troubled" về cùng một từ "troubl" không đúng chính tả và không xuất hiện trong từ điển tiếng Anh.
- NLTK toolkit cũng cung cấp các công cụ stemmer này

Lemmatization và Stemming

Sample text: Such an analysis can reveal features that are not easily visible from the variations in the individual genes and can lead to a picture of expression that is more biologically transparent and accessible to interpretation

Lovins stemmer: such an analys can reve featur that ar not eas vis from th vari in th individu gen and can lead to a pictur of expres that is mor biolog transpar and acces to interpres

Porter stemmer: such an analysi can reveal featur that ar not easili visibl from the variat in the individu gene and can lead to a pictur of express that is more biolog transpar and access to interpret

Paice stemmer: such an analys can rev feat that are not easy vis from the vary in the individ gen and can lead to a pict of express that is mor biolog transp and access to interpret

Figure 2.8: A comparison of three stemming algorithms on a sample text.

https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

Lemmatization và Stemming

Implementation example:

```
from nltk.stem import PorterStemmer
words = ["operate", "operating", "operates", "operation",
"operative", "operatives", "operational"]

ps = PorterStemmer()

for token in words:
    print (ps.stem(token))
```

Output:

```
oper
oper
oper
oper
oper
oper
oper
```

https://towardsdatascience.com/building-blocks-text-pre-processing-641cae8ba3bf

LNK 55

Stemming Words

1

```
from nltk.stem import PorterStemmer
from nltk.stem import LancasterStemmer
```

2

```
#create an object of class PorterStemmer
porter = PorterStemmer()
lancaster=LancasterStemmer()
#proide a word to be stemmed
print("Porter Stemmer")
print(porter.stem("cats"))
print(porter.stem("trouble"))
print(porter.stem("troubling"))
print(porter.stem("troubled"))
print("Lancaster Stemmer")
print(lancaster.stem("cats"))
print(lancaster.stem("trouble"))
print(lancaster.stem("troubling"))
print(lancaster.stem("troubled"))
```

3

```
Porter Stemmer
cat
troubl
troubl
troubl
Lancaster Stemmer
cat
troubl
troubl
troubl
```

https://www.datacamp.com/community/tutorials/stemming-lemmatization-python

LNK 56

Stemming Sentences

- We need to stem each word in the sentence and return a combined sentence.
- To separate the sentence into words, you can use tokenizer.
- The *NLTK tokenizer* separates the sentence into words as follows.

```
from nltk.tokenize import sent_tokenize, word_tokenize
def stemSentence(sentence):
    token_words=word_tokenize(sentence)
    token_words
    stem_sentence=[]
    for word in token_words:
        stem_sentence.append(porter.stem(word))
        stem_sentence.append(" ")
    return "".join(stem_sentence)

x=stemSentence(sentence)
print(x)
```

LNK_57

Stemming Sentences

sentence="Pythoners are very intelligent and work very pythonly and now they are pythoning their way to success."

porter.stem(sentence)

```
from nltk.tokenize import sent_tokenize, word_tokenize
def stemSentence(sentence):
    token_words=word_tokenize(sentence)
    token_words
    stem_sentence=[]
    for word in token_words:
        stem_sentence.append(porter.stem(word))
        stem_sentence.append(" ")
    return "".join(stem_sentence)

x=stemSentence(sentence)
print(x)
```

```
python are veri intellig and work veri pythonli and now they are python their way to success .
```

LNK_58

Stemming a document

You can write your own function that can stem documents.

1. Take a document as the input.
2. Read the document line by line
3. Tokenize the line
4. Stem the words
5. Output the stemmed words (print on screen or write to a file)
6. Repeat step 2 to step 5 until it is to the end of the document.

Lemmatization hay Stemming ?

It depends on the application

- Stemming and Lemmatization both generate the root form of the inflected words. *The difference is that stem might not be an actual word whereas, lemma is an actual language word.*
- Stemming follows an algorithm with steps to perform on the words which makes it faster. Whereas, in lemmatization, you used WordNet corpus and a corpus for stop words as well to produce lemma which makes it slower than stemming. You also had to define a parts-of-speech to obtain the correct lemma.

Stopwords

Wikipedia:

"In computing, stop words are words which are filtered out before processing of natural language data (text). Stop words are generally the most common words in a language; there is no single universal list of stop words used by all natural language processing tools, and indeed not all tools even use such a list. Some tools avoid removing stop words to support phrase search."

Removing stop words with NLTK

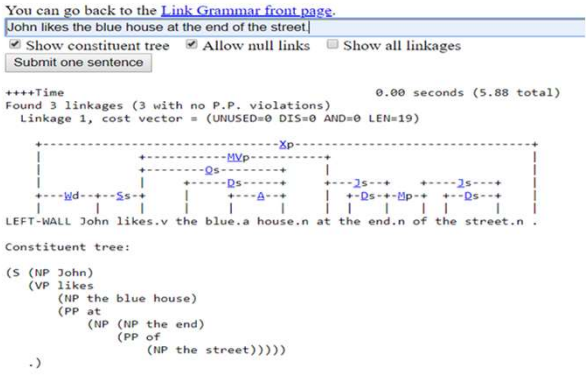
NLTK has a list of stopwords stored in 16 different languages.

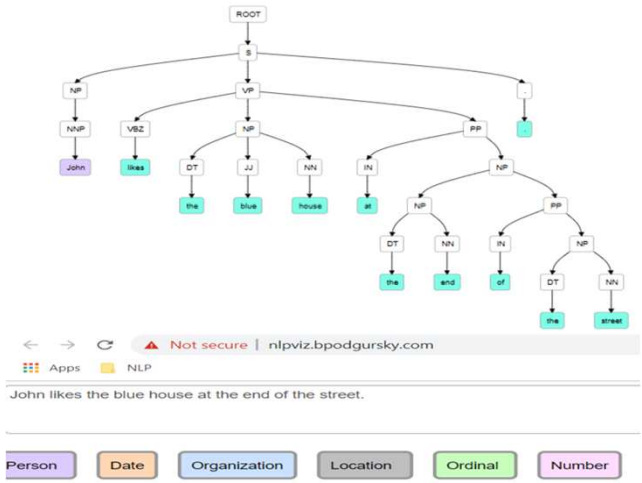
https://www.geeksforgeeks.org/removing-stop-words-nltk-python/

https://medium.com/@makcedward/nlp-pipeline-stop-words-part-5-d6770df8a936

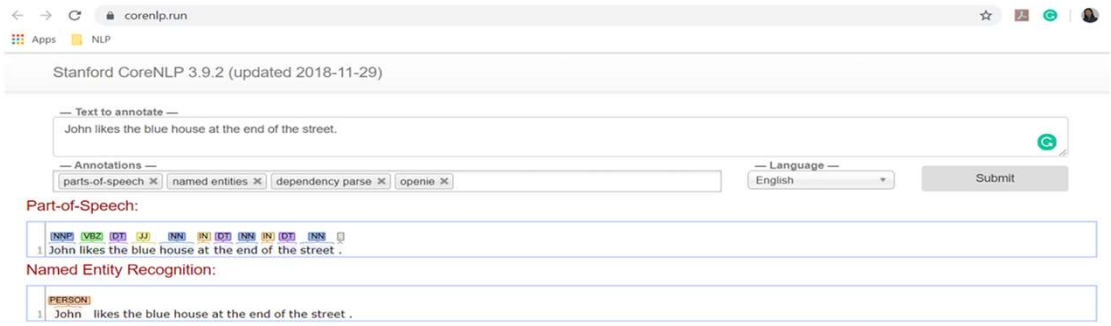
Parsing

Building the syntactic tree of a sentence



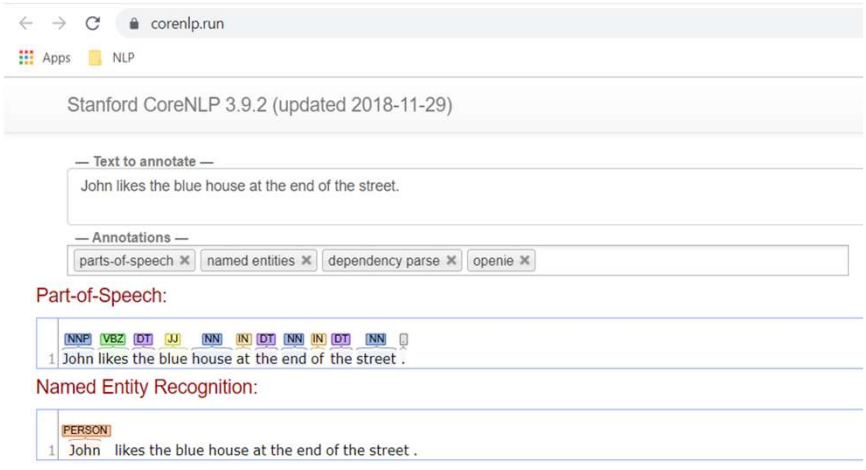


POS tagging



LNK_63

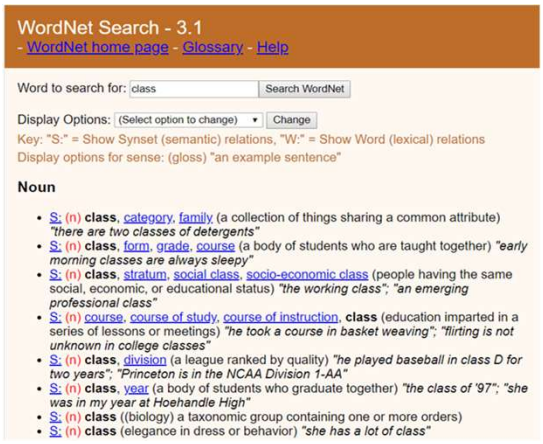
Named-entity recognition



LNK_64

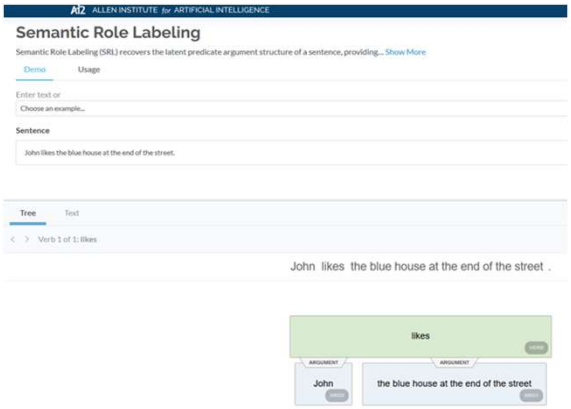
Word sense disambiguation

➢Figuring out the exact meaning of a word or an entity



Semantic role labeling

- Extracting subject-predicate-object triples from a sentence
- Semantic role labeling also called shallow semantic parsing or slot-filling



Nội dung Chương

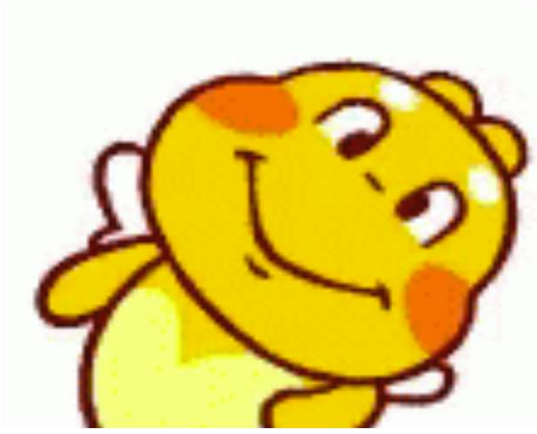
1. Giới thiệu tổng quan NLP
2. Các mức độ ngôn ngữ
3. Tài nguyên ngôn ngữ
4. Hướng tiếp cận
5. **Phương pháp đánh giá trong NLP**
6. Vì sao NLP khó
7. Công cụ tách từ tiếng Việtase study

Đánh giá trong NLP

- Đo lường hay đánh giá việc thực thi của một mô hình đã xây dựng trong NLP là cực kỳ quan trọng, và cần lựa chọn đúng chỉ số đánh giá.
- NLP sử dụng 2 hình thức đánh giá:
  - Đánh giá điểm cuối (end-to-end) hay đánh giá bên ngoài (extrinsic evaluation) và
  - Đánh giá bên trong (intrinsic evaluation) hay đánh giá nội tại.
    - Precision, Recall
    - F1-score
    - Accuracy
    - MRR (Mean Reciprocal Rank)
    - MAP (Mean Average Precision)
    - BLEU (Bilingual Evaluation Understudy)
    - ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

Nội dung Chương

1. Giới thiệu tổng quan NLP
2. Các mức độ ngôn ngữ
3. Tài nguyên ngôn ngữ
4. Hướng tiếp cận
5. Phương pháp đánh giá trong NLP
6. **Vì sao NLP khó**
7. Công cụ tách từ tiếng Việt



Paraphrasing

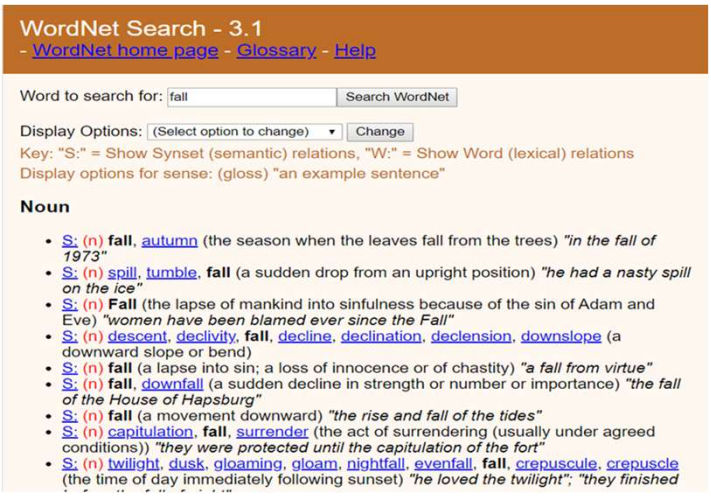
Các từ/câu khác nhau có thể diễn đạt cùng ý nghĩa (meaning)

- Season of the year
  - Fall
  - Autumn
- Book delivery time
  - When will my book arrive?
  - When will I receive my book?

Mơ hồ

Mơ hồ ngữ nghĩa (ambiguity):

Một từ hoặc câu có thể có nhiều nghĩa khác nhau



Phonetics and Phonology

- Phonology ambiguity: nghĩa là các từ có âm giống nhau nhưng nghĩa khác nhau
- Ví dụ:
  - there-their
  - sea-see
  - ice Cream - I scream
  - I don't know - I dont. No!
  - **Kiến bò đĩa thịt bò/ Ruồi đậu mâm xôi đậu.**

Syntax and ambiguity

I saw the man with a telescope.

→ Who had the telescope?

Syntax and ambiguity

- Lexical ambiguity
  - Outside of a dog, a book is a man's best friend. Inside of a dog, it's too dark to read.
- Structural ambiguity
  - One morning I shot an elephant in my pajamas. How he got into my pajamas I'll never know.

- Buffalo buffalo buffalo buffalo buffalo buffalo buffalo buffalo
- The complex houses married and single soldiers and their families

LNK_73

Semantics

The astronomer loves the star.

https://en.wikipedia.org/wiki/Star#/media/File:Starsinthesky.jpg

http://www.businessnewsdaily.com/2023-celebrity-hiring.html

LNK_74

Nội dung Chương

1. Giới thiệu tổng quan NLP
2. Các mức độ ngôn ngữ
3. Tài nguyên ngôn ngữ
4. Hướng tiếp cận
5. Phương pháp đánh giá trong NLP
6. Vì sao NLP khó
7. **Công cụ tách từ tiếng Việt**

75

Các công cụ tách từ tiếng Việt

Tiếng Việt:

- Tiếng Việt thuộc nhóm ngôn ngữ Việt-Mường, một nhóm nhỏ trong nhóm Mon-Khmer, và Mon-Khmer là một thành viên trong họ ngôn ngữ Auso-Asiatic.
- Ngôn ngữ tiếng Việt tuân theo cấu trúc SVO, không có hình thái học ngôn ngữ, không sử dụng mạo từ và không dùng thì bị động.
- Các mối quan hệ ngữ pháp được thể hiện bằng việc sử dụng các trợ từ (auxiliary word) và trật tự từ.
  - Ví dụ, thì quá khứ của động từ "đi" ở tiếng Việt là "đã đi", số nhiều của từ "ngôi nhà" là "nhiều ngôi nhà".

76

Các công cụ tách từ tiếng Việt

Tiếng Việt:

- Không giống như tiếng Anh, khoảng trắng không được sử dụng để phân tách các từ trong tiếng Việt.
- Phần có ý nghĩa nhỏ nhất trong tiếng Việt là "âm tiết" (syllabel).
- Một từ trong tiếng Việt có thể gồm nhiều âm tiết cách nhau bằng khoảng trắng.

Các công cụ tách từ tiếng Việt

Tên công cụ	Địa chỉ download
JVnSegmenter	http://jvnsegmenter.sourceforge.net/
DongDu	https://github.com/rockkhuya/DongDu
Pyvi	https://pypi.org/project/pyvi/
RDRsegmenter	https://github.com/datquocnguyen/RDRsegmenter
UETsegmenter	https://github.com/phongnt570/UETsegmenter
UITws	https://github.com/ngannlt/UITws-v1
Underthesea	https://github.com/undertheseanlp/underthesea
Vitk	https://github.com/phuonglh/vn.vitk
VnCoreNLP	https://github.com/vncorenlp/VnCoreNLP
vnTokenizer	https://vlsp.hpda.vn/demo/?page=resources