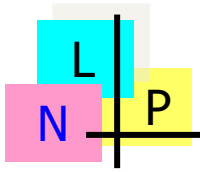
INDUSTRIAL UNIVERSITY OF HOCHIMINH CITY

Trường Đại học Công nghiệp Thành phố Hồ Chí Minh

XỬ LÝ NGÔN NGỮ TỰ NHIÊN

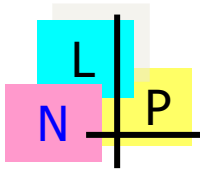
Natural Language Processing - (NLP)

Giảng viên: Tiến sĩ Bùi Thanh Hùng
Bộ môn Khoa học dữ liệu
Khoa Công nghệ thông tin
Đại học Công nghiệp TP HCM
Email: buithanhhung@iuh.edu.vn
Website: https://sites.google.com/site/hungthanhbui1980/



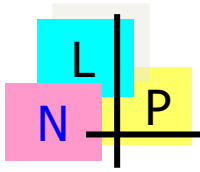
Outline

- Overview of the field
  - What is Natural Language Processing?
  - NLP applications
  - Aspects of language processing
  - Why NLP is difficult?
- The NLP Research Community



What is Natural Language Processing?

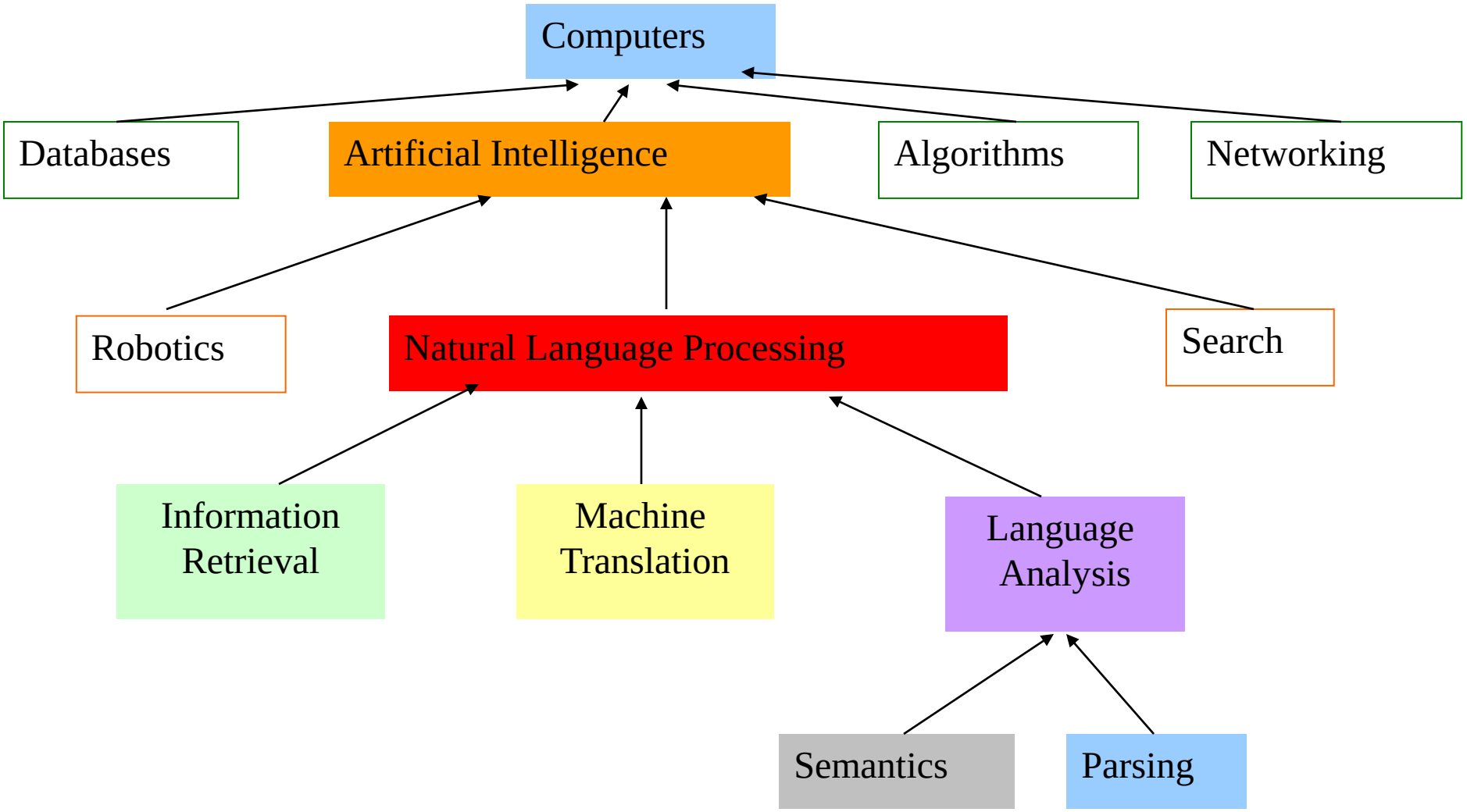
- Natural Language Processing
  - Process information contained in natural language text.
  - Also known as Computational Linguistics (CL), Human Language Technology (HLT), Natural Language Engineering (NLE)

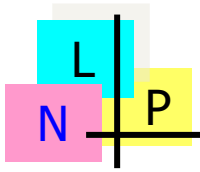


NLP applications

- Text Categorization
  - Classify documents by topics, language, author, spam filtering, information retrieval (relevant, not relevant), sentiment classification (positive, negative)
- Spelling & Grammar Corrections
- Information Extraction
- Speech Recognition
- Information Retrieval
  - Synonym Generation
- Summarization
- Machine Translation
- Question Answering
- Dialog Systems
  - Language generation

Where does it fit in the CS taxonomy?





Aspects of language processing

- Word, lexicon: lexical analysis
  - Morphology, word segmentation
- Syntax
  - Sentence structure, phrase, grammar, …
- Semantics
  - Meaning
  - Execute commands
- Discourse analysis
  - Meaning of a text
  - Relationship between sentences (e.g. anaphora)

Dependency Parsing

Raw sentence

He reckons the current account deficit will narrow to only 1.8 billion in September.

Part-of-speech tagging

POS-tagged sentence

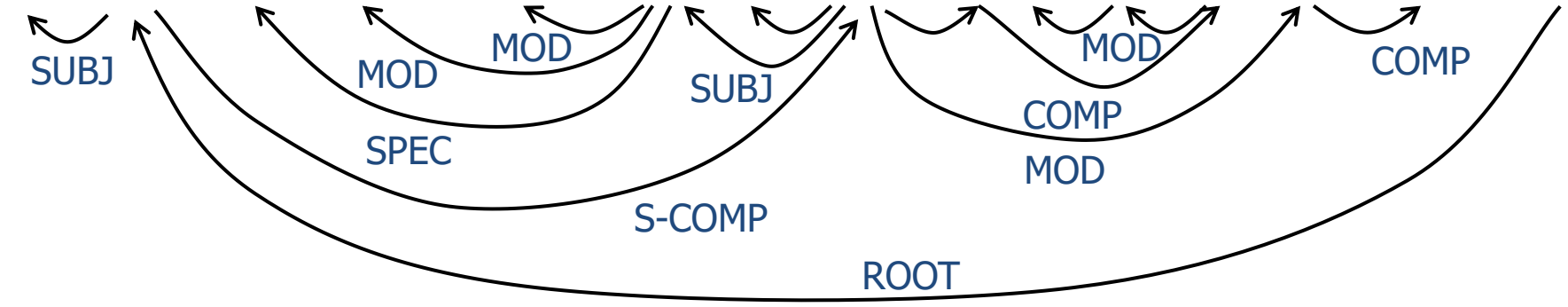
He reckons the current account deficit will narrow to only 1.8 billion in September.

PRP VBZ DT JJ NN NN MD VB TO RB CD CD IN NNP .

Word dependency parsing

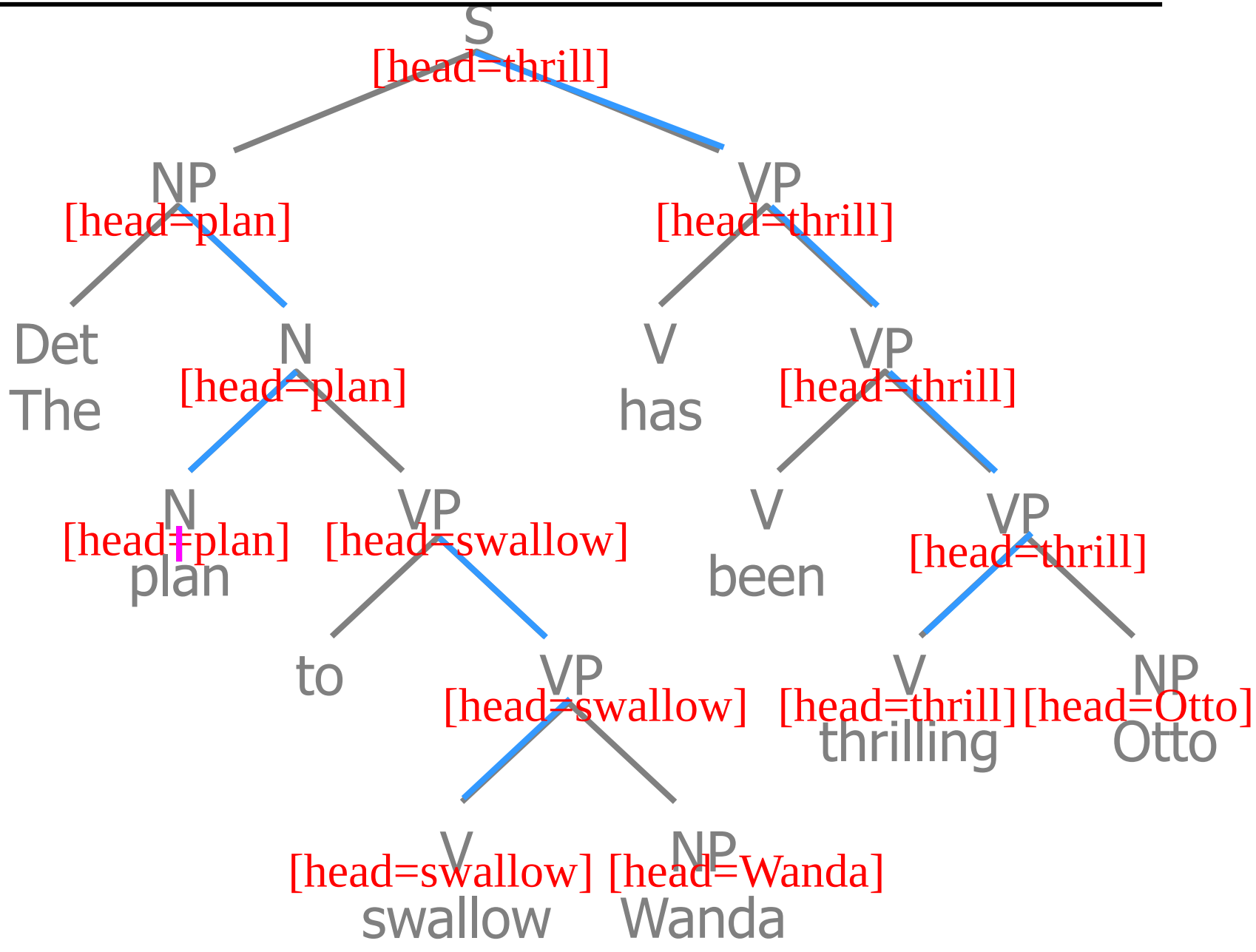
Word dependency parsed sentence

He reckons the current account deficit will narrow to only 1.8 billion in September .



Dependency Trees

1. Assign heads

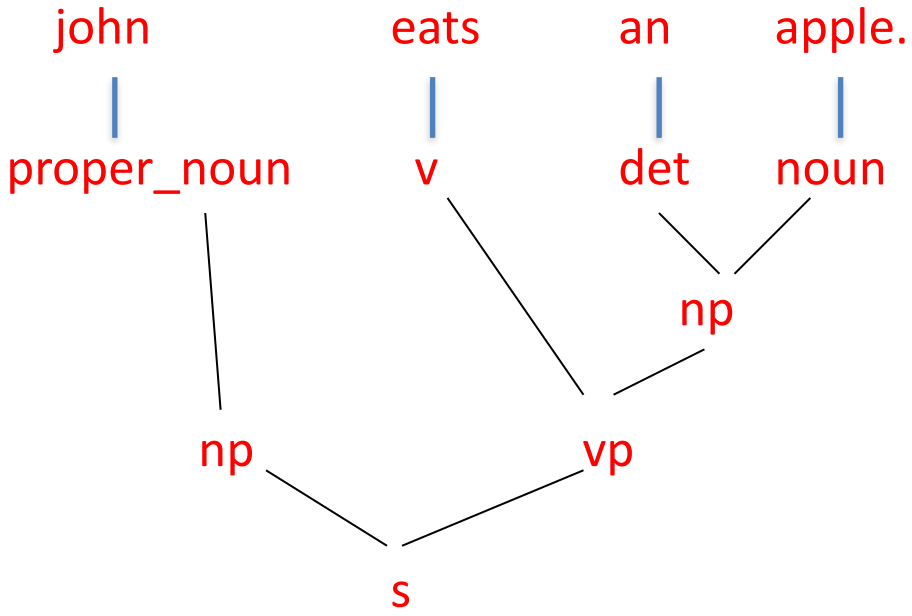


Parsing (in Definite Clause Grammars)

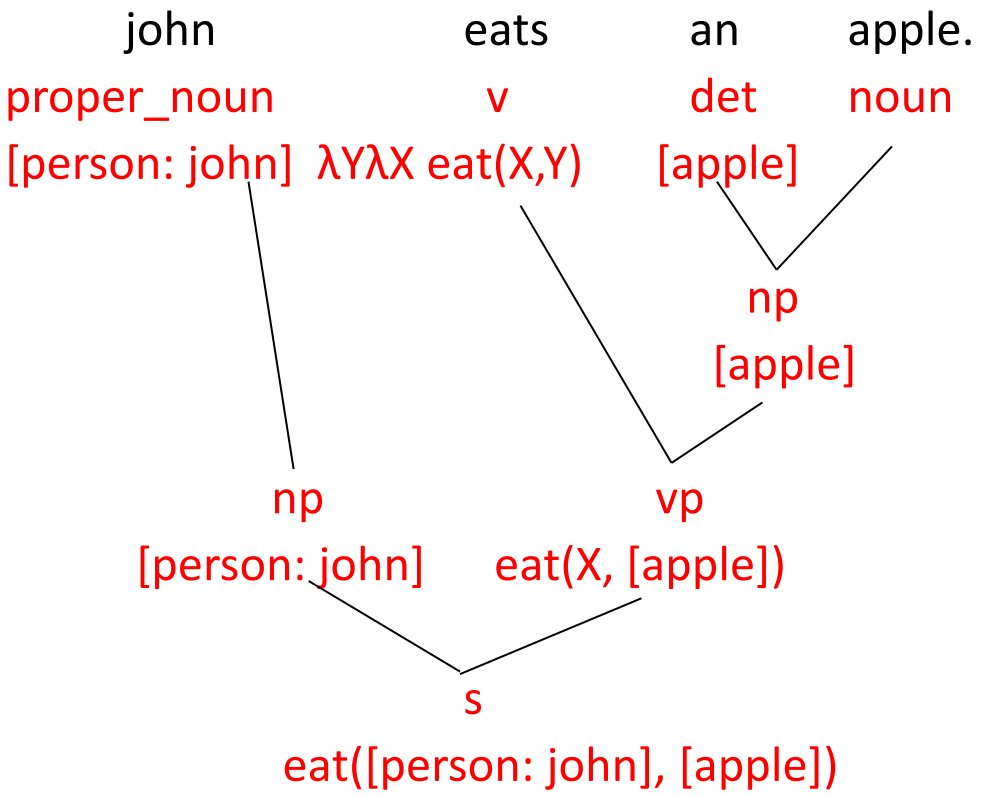
s --> np, vp
np --> det, noun
np --> proper_noun
vp --> v, np
vp --> v.

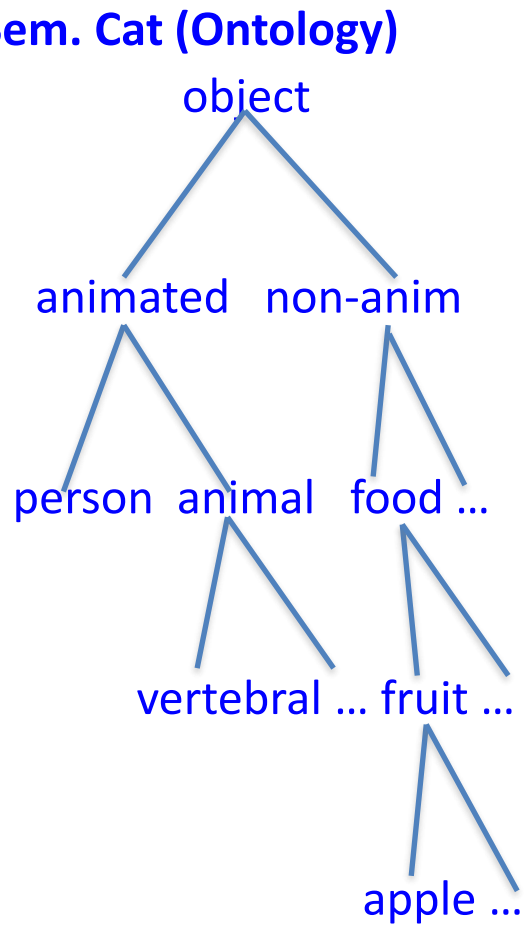
det -->[a]. det --> [an].
det --> [the].
noun --> [apple].
noun --> [orange].
proper_noun --> [john].
proper_noun --> [mary].
v --> [eats].
v --> [loves].

Eg.



Semantic analysis





Parsing & semantic analysis

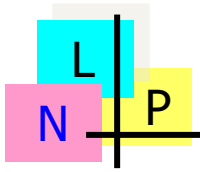
- Rules: syntactic rules or semantic rules
  - What component can be combined with what component?
  - What is the result of the combination?
- Categories
  - Syntactic categories: Verb, Noun, …
  - Semantic categories: Person, Fruit, Apple, …
- Analyses
  - Recognize the category of an element
  - See how different elements can be combined into a sentence
  - Problem: The choice is often not unique

Discourse analysis

- Anaphora

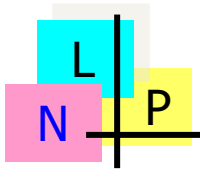
  He hits the car with a stone. It bounces back.

- Understanding a text
  - Who/when/where/what … are involved in an event?
  - How to connect the semantic representations of different sentences?
  - What is the cause of an event and what is the consequence of an action?
  - …



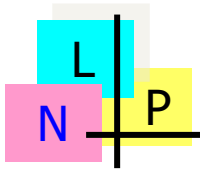
NLP

Cleanup, Tokenization	Information Retrieval and Extraction (IR)	Machine Translation
Stemming	Relationship Extraction	Automatic Summarization/ Paraphracing
Lemmatization	Named Entity Recognition (NER)	Natural Language Generation
Part of Speech Tagging	Sentiment Analysis/Sentance Boundary Disambiguation	Reasoning over Knowledge Based
Query Expansion	Word sense and Disambiguation	Question Answering System
Parsing	Text Similarity	Dialog System
Topic Segmentation and Recognition	Coreference Resolution	Image Captioning & other Multimodel Tasks
Morphological Degmentation (Word/Sentences)	Discourse Analysis	



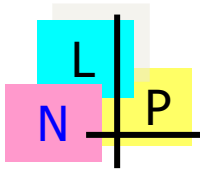
Why NLP is difficult

- A NLP system needs to answer the question "who did what to whom"
- **Language is ambiguous**
  - At all levels: lexical, phrase, semantic
  - Iraqi Head Seeks Arms
    - Word sense is ambiguous (head, arms)
  - Stolen Painting Found by Tree
    - Thematic role is ambiguous: tree is agent or location?
  - Ban on Nude Dancing on Governor's Desk
    - Syntactic structure (attachment) is ambiguous: is the ban or the dancing on the desk?
  - Hospitals Are Sued by 7 Foot Doctors
    - Semantics is ambiguous : what is 7 foot?



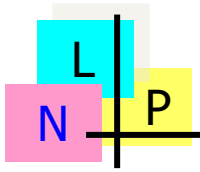
Why NLP is difficult

- Language is flexible
  - New words, new meanings
  - Different meanings in different contexts
- Language is subtle
  - He arrived at the lecture
  - He chuckled at the lecture
  - He chuckled his way through the lecture
  - **He arrived his way through the lecture
- Language is complex!



Why NLP is difficult

- MANY hidden variables
  - Knowledge about the world
  - Knowledge about the context
  - Knowledge about human communication techniques
    - *Can you tell me the time?*
- Problem of scale
  - Many (infinite?) possible words, meanings, context
- Problem of sparsity
  - Very difficult to do statistical analysis, most things (words, concepts) are never seen before
- Long range correlations

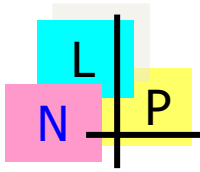


Why NLP is difficult

- Key problems:
  - Representation of *meaning*
  - Language presupposes knowledge about the world
  - Language only reflects the surface of meaning
  - Language presupposes communication between people

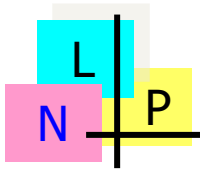
Meaning

- What is meaning?
  - Physical referent in the real world
  - Semantic concepts, characterized also by relations.
- How do we represent and use meaning
  - I am *Italian*
    - *From lexical database* (WordNet)
    - *Italian =a native or inhabitant of Italy → Italy = republic in southern Europe [..]*
  - I am Italian
    - Who is "I"?
  - I know she is Italian/I think she is Italian
    - How do we represent "I know" and "I think"
    - Does this mean that I is Italian? What does it say about the "I" and about the person speaking?
  - I thought she was Italian
    - How do we represent tenses?



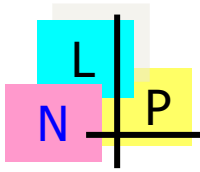
The NLP Research Community

- Papers
  - ACL Anthology has nearly everything, free!
    - Over 20,000 papers!
    - Free-text searchable
      - Great way to learn about current research on a topic
      - New search interfaces currently available in beta
        - » Find recent or highly cited work; follow citations
    - Used as a dataset by various projects
      - Analyzing the text of the papers (e.g., parsing it)
      - Extracting a graph of papers, authors, and institutions (Who wrote what? Who works where? What cites what?)



The NLP Research Community

- Conferences
  - Most work in NLP is published as 8-page conference papers with 3 double-blind reviewers.
  - Main annual conferences: ACL, EMNLP, NAACL
    - Also EACL, IJCNLP, COLING
    - + various specialized conferences and workshops
  - Big events, and growing fast! ACL 2014:
    - About 1000 attendees
    - 572 full-length papers submitted (146 accepted)
    - 551 short papers submitted (139 accepted)
    - 16 workshops on various topics



The NLP Research Community

NLP Journals

Computational Linguistics

Journal of Natural Language Engineering (JLNE)

Machine Translation

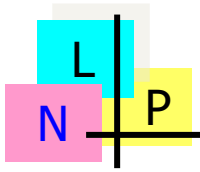
Natural Language and Linguistic Theory

Journal of Natural Language Processing

…

The NLP Research Community

- Institutions
  - Universities: Many have NLP faculty
    - Several "big players" with many faculty
    - Some of them also have good linguistics, cognitive science, machine learning, AI
  - Companies:
    - Old days: AT&T Bell Labs, IBM
    - Now: Google, Microsoft, IBM, many startups …
      - Speech: Nuance, …
      - Machine translation: Language Weaver, Systran, …
      - Many niche markets – online reviews, medical transcription, news summarization, legal search and discovery …



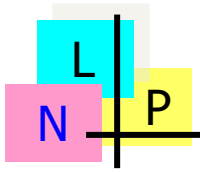
The NLP Research Community

NLP Research Centers

AT&T Labs - Research
BBN Systems and Technologies Corporation
DFKI (German research center for AI)
General Electric R&D
IRST, Italy
IBM T.J. Watson Research, NY
Lucent Technologies Bell Labs, Murray Hill, NJ
Microsoft Research, Redmond, WA
MITRE
NEC Corporation
SRI International, Menlo Park, CA
SRI International, Cambridge, UK
Xerox, Palo Alto, CA
XRCE, Grenoble, France
Google, Microsoft, Facebook, Amazon, …

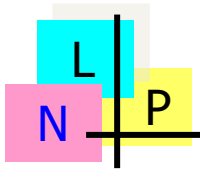
The NLP Research Community

- Standard tasks
  - If you want people to work on your problem, make it easy for them to get started and to measure their progress. Provide:
    - Test data, for evaluating the final systems
    - Development data, for measuring whether a change to the system helps, and for tuning parameters
    - An evaluation metric (formula for measuring how well a system does on the dev or test data)
    - A program for computing the evaluation metric
    - Labeled training data and other data resources
    - A prize? – with clear rules on what data can be used



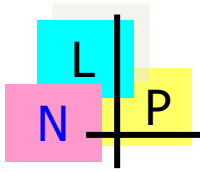
The NLP Research Community

- Software
  - Lots of people distribute code for these tasks
    - Or you can email a paper's authors to ask for their code
  - Some lists of software, but no central site ☹
  - Some end-to-end pipelines for text analysis
    - "One-stop shopping"
    - Cleanup/tokenization + morphology + tagging + parsing + …
    - NLTK is easy for beginners and has a free book (intersession?)
    - GATE has been around for a long time and has a bunch of modules



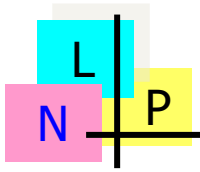
The NLP Research Community

- Software
  - To find good or popular tools:
    - Search current papers, ask around, use the web
  - Still, often hard to identify the **best** tool for your job:
    - Produces appropriate, sufficiently detailed output?
    - Accurate? (on the measure you care about)
    - Robust? (accurate on your data, not just theirs)
    - Fast?
    - Easy and flexible to use? Nice file formats, command line options, visualization?
    - Trainable for new data and languages? How slow is training?
    - Open-source and easy to extend?



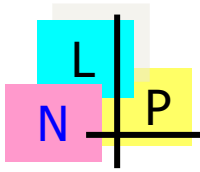
The NLP Research Community

- Datasets
  - Raw text or speech corpora
    - Or just their n-gram counts, for super-big corpora
    - Various languages and genres
    - Usually there's some metadata (each document's date, author, etc.)
    - Sometimes ∃ licensing restrictions (proprietary or copyright data)
  - Text or speech with manual or automatic annotations
    - What kind of annotations? That's the rest of this lecture …
    - May include translations into other languages
  - Words and their relationships
    - Morphological, semantic, translational, evolutionary
  - Grammars
  - World Atlas of Linguistic Structures
  - Parameters of statistical models (e.g., grammar weights)



The NLP Research Community

- Datasets
  - Read papers to find out what datasets others are using
    - Linguistic Data Consortium (searchable) hosts many large datasets
    - Many projects and competitions post data on their websites
    - But sometimes you have to email the author for a copy
  - CORPORA mailing list is also good place to ask around
  - LREC Conference publishes papers about new datasets & metrics
  - Amazon Mechanical Turk – pay humans (very cheaply) to annotate your data or to correct automatic annotations
    - Old task, new domain: Annotate parses etc. on *your* kind of data
    - New task: Annotate something new that you want your system to find
    - Auxiliary task: Annotate something new that your system may benefit from finding (e.g., annotate subjunctive mood to improve translation)
  - Can you make annotation so much fun or so worthwhile that they'll do it for free?

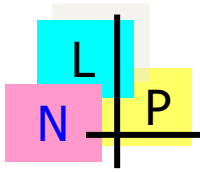


The NLP Research Community

Datasets

Một số nguồn để tìm dataset về machine learning, data science, AI

1. Google Datasets:
Link : https://datasetsearch.research.google.com/
2. Papers with Code Datasets.
Link : https://paperswithcode.com/datasets
3. Kaggle Dataset
Link: https://www.kaggle.com/datasets
4. Big Bag NLP Datasets
Link: https://index.quantumstat.com/#/
5. Hugging Face Datasets
Link: https://huggingface.co/dataset
6. UCI Machine Learning
Link: https://archive.ics.uci.edu/ml/index.php



The NLP Research Community

Datasets

Một số nguồn để tìm dataset về machine learning, data science, AI.

7. Amazin Datasets (Open Data on AWS)

Link: https:/aws.amazon.com/opendata/

8. Awesome Public Datasets

Link: https://github.com/awesomedata/awesome-public-datasets

9. Azure public datasets

Link: https://docs.microsoft.com/.../azure-sql/public-data-sets

10. Carnegie Mellon University

Link: https://guides.library.cmu.edu/az.php

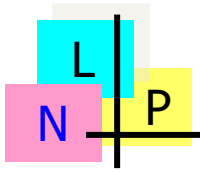
11. .gov Datasets

Link: https://data.gov.au/

https://data.gov.in/

https://data.gov.sg/

https://data.europa.eu/data/datasets?locale=en&minScoring=0



The NLP Research Community

Datasets

Một số nguồn để tìm dataset về machine learning, data science, AI.

7. Amazin Datasets (Open Data on AWS)

Link: https:/aws.amazon.com/opendata/

8. Awesome Public Datasets

Link: https://github.com/awesomedata/awesome-public-datasets

9. Azure public datasets

Link: https://docs.microsoft.com/.../azure-sql/public-data-sets

10. Carnegie Mellon University

Link: https://guides.library.cmu.edu/az.php

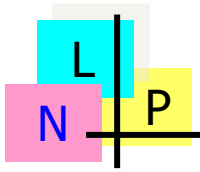
11. .gov Datasets

Link: https://data.gov.au/

https://data.gov.in/

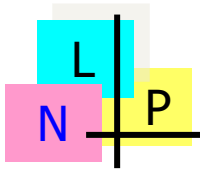
https://data.gov.sg/

https://data.europa.eu/data/datasets?locale=en&minScoring=0



The NLP Research Community

- Standard data formats
  - Often just simple *ad hoc* text-file formats
    - Documented in a README; easily read with scripts
  - Some standards:
    - Unicode – strings in any language (see ICU toolkit)
    - PCM (.wav, .aiff) – uncompressed audio
      - BWF and AUP extend w/metadata; also many compressed formats
    - XML – documents with embedded annotations
    - Text Encoding Initiative – faithful digital representations of printed text
    - Protocol Buffers, JSON – structured data
    - UIMA – "unstructured information management"; Watson uses it
  - Standoff markup: raw text in one file, annotations in other files ("∃ noun phrase from byte 378—392")
    - Annotations can be independently contributed & distributed



The NLP Research Community

- Survey articles
  - May help you get oriented in a new area
  - Synthesis Lectures on Human Language Technologies
  - Handbook of Natural Language Processing
  - Oxford Handbook of Computational Linguistics
  - Foundations & Trends in Machine Learning
  - ACM Computing Surveys?
  - Online tutorial papers
  - Slides from tutorials at conferences
  - Textbooks

The NLP Research Community

- Vietnam

Jaist: GS Nguyễn Lê Minh

Trường Đại học Công nghệ, ĐHQGHN

Vin University

Đại học KHTN

Đại học Bách khoa

Đại học CNTT

Học viện Bưu chính viễn thông

Đại học Kyoto: TS Phạm Quang Nhật Minh

Đại học Tôn Đức Thắng, Đại học Kỹ thuật CN, Đại học Hà Nội

…

The NLP Research Community

- Toolkits

Tsujii Lab-Tokyo, Japan: http://www.nactem.ac.uk/tsujii/
Stanford Lab, America: http://nlp.stanford.edu/
Matsumoto Lab-NAIST, Japan: http://cl.naist.jp/en/
NLTK Toolkits: http://www.nltk.org/
Open NLP: http://opennlp.sourceforge.net/projects.html
NLP Toolkits: http://www.phontron.com/nlptools.php
Kyoto Lab: http://nlp.ist.i.kyoto-u.ac.jp/EN/
Google NLP research: http://research.google.com/pubs/NaturalLanguageProcessing.html

VLSP project: http://vlsp.vietlp.org:8080/demo/?&lang=en
Nguyễn Lê Minh: http://www.jaist.ac.jp/~nguyenml/
Lưu Văn Hải, Nguyễn Tuấn Hải, Japan: http://viet.jnlp.org/

The NLP Research Community

https://github.com/undertheseanlp/NLP-Vietnamese-progress

- Sentence Boundary Disambiguation / Language Detection / Text Normalization / Spelling Correction
- Word Segmentation / Part-of-Speech Tagging / Chunking / Parsing
- Text Classification / Sentiment Analysis / Word Embeddings
- Named Entity Recognition / Relationship Extraction / Event Extraction / Information Extraction / Keyword Extraction
- Coreference Resolution / Slot Filling / Entity Linking
- Semantics / Semantic Role Labeling / Paraphrase Identification / Natural Language Inference
- Machine Translation / Automatic Summarization
- Knowledge Representation and Reasoning
- Dialog Systems and Chatbots / Language Generation / Question Answering
- Automatic Speech Recognition / Text To Speech / Speech Classification / Speech
- Optical Text Recognition / Image Captioning
- Resources

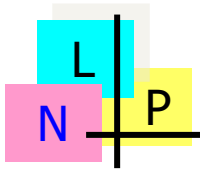
The NLP Research Community

https://github.com/undertheseanlp/NLP-Vietnamese-progress

Named Entity Recognition

Without gold POS and chunking tags

Model	F1	Paper/Source	Code
PhoBERT-large	94.7	Nguyen et al. '20	Official
PhoBERT-base	93.6	Nguyen et al. '20	Official
VnCoreNLP used ETNLP embeddings	91.30	Nguyen et al. NAACL'18	Official
VNER Attentive Neural Network	90.37	Dong et al. '18	
vietner CRF (ngrams + word shapes + cluster + w2v)	90.03	Pham CICLing'18	Official
VnCoreNLP dynamic feature induction model	88.55	Nguyen et al. NAACL'18	Official



Summary

- Overview of the field
  - What is Natural Language Processing?
  - NLP applications
  - Aspects of language processing
  - Why NLP is difficult?
- The NLP Research Community