XỬ LÝ NGÔN NGỮ TỰ NHIÊN

CHƯƠNG 1 – GIỚI THIỆU

NGUYỄN TRỌNG CHỈNH

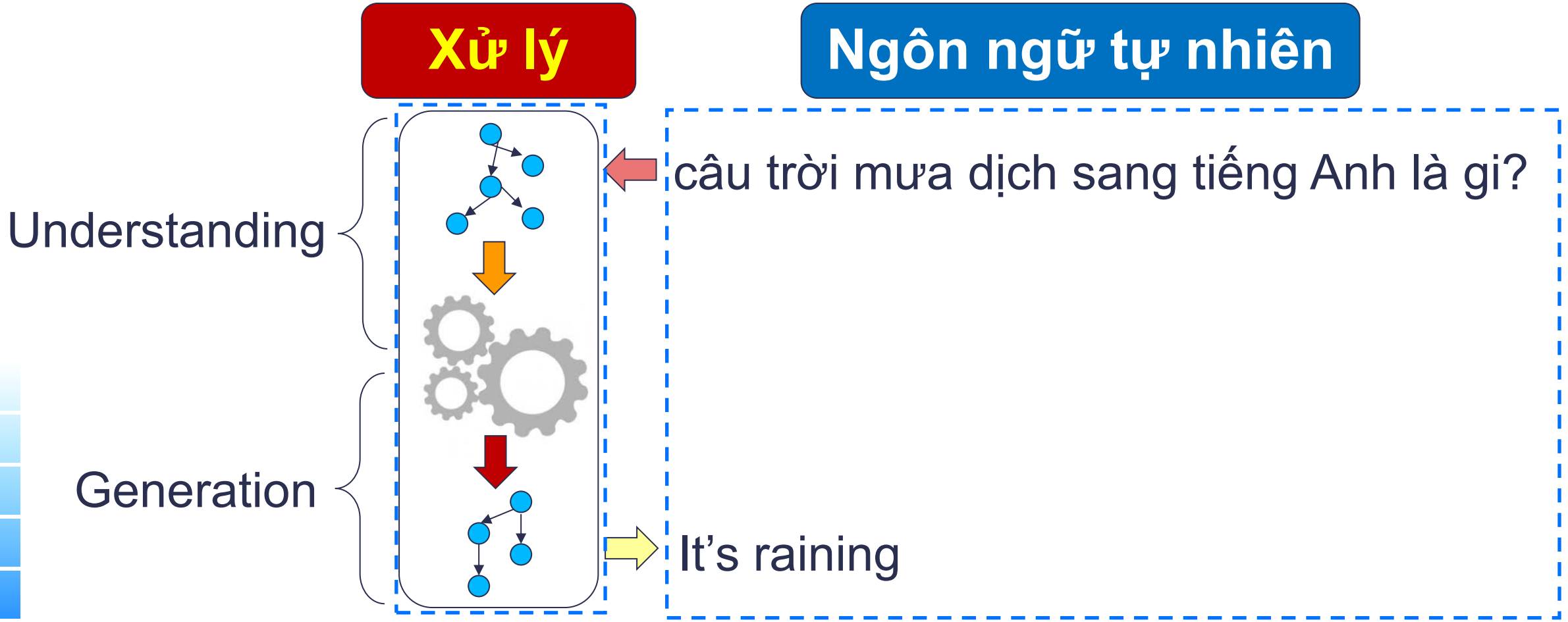
TRÌNH BÀY

1. XỬ LÝ NGÔN NGỮ TỰ NHIÊN LÀ GÌ?
2. CÁC CƠ SỞ
3. MỘT SỐ ỨNG DỤNG

XLNNTN LÀ GÌ

XLNNTN LÀ GÌ?

Xử lý ngôn ngữ tự nhiên (Natural Language Processing – NLP)



XLNNTN LÀ GÌ?

NGÔN NGỮ TỰ NHIÊN (natural language)	NGÔN NGỮ NHÂN TẠO (artificial language)
- Hình thành tự phát. - Quy tắc phức tạp, nhiều ngoại lệ - Mơ hồ (ambiguity). - Sử dụng để giao tiếp, lưu truyền văn hóa, kiến thức và kinh nghiệm.	- Xây dựng có chủ đích. - Quy tắc đơn giản, ít ngoại lệ. - Rõ ràng. - Sử dụng cho từng mục đích chuyên biệt

XLNNTN LÀ GÌ?

XỬ LÝ NGÔN NGỮ TỰ NHIÊN (NLP)	NGÔN NGỮ HỌC TÍNH TOÁN (Computational Linguistics)
- Mục đích: xây dựng ứng dụng xử lý, phân tích, tạo sinh ngôn ngữ. - Giải quyết các bài toán thực tế bằng cách nghiên cứu và áp dụng các mô hình tính toán. - Nghiên cứu phát triển các ứng dụng, hệ thống thực tế.	- Mục đích: hiểu ngôn ngữ tự nhiên. - Phân tích các hiện tượng trong ngôn ngữ bằng các phương pháp tính toán. - Nghiên cứu lý thuyết, mô hình biểu diễn ngôn ngữ tự nhiên.

XLNNTN LÀ GÌ?

Vài nét về lịch sử của XLNNTN

1. 1950s. Dịch máy sử dụng hệ luật (rule-based).
2. 1960s-1970s. Văn phạm và các phương pháp hình thức.
    - Văn phạm tạo sinh (generative grammar) của Chomsky.
    - Phân tích cú pháp.
    - Chương trình SHRDLU, dùng ngôn ngữ tự nhiên để điều khiển vật thể được hiển thị trên màn hình.
    - Chương trình ELIZA.

XLNNTN LÀ GÌ?

Vài nét về lịch sử của XLNNTN (tt)

3. 1980s. Các nghiên cứu về biểu diễn tri thức và suy luận.
    - Các thuật toán phân tích cú pháp dựa trên hệ luật.
    - Mạng ngữ nghĩa.
    - Biểu diễn ngữ nghĩa theo logic hình thức
4. 1990s. Mô hình xác suất với ngữ liệu.
    - Hidden Markov, n-gram, phân tích cú pháp có xác suất.
    - Ngữ liệu Penn Treebank.

XLNNTN LÀ GÌ?

Vài nét về lịch sử của XLNNTN (tt)

5. 2000s. Học máy xác suất.
    - SVM, MEMM, CRF.
    - Vấn đề nhận dạng thực thể (NER), rút trích thông tin (IE).
6. 2010s. Mạng nơron.
    - Word embeddings
    - RNN, LSTM, CNN, BERT, GPT.
    - Sequence to sequence

XLNNTN LÀ GÌ?

Vài nét về lịch sử của XLNNTN (tt)

7. 2020s. Các mô hình ngôn ngữ lớn.
    - Few-shot learning, zero-shot learning.
    - Đa phương sắc thái (multi-modal).

XLNNTN LÀ GÌ?

Một số thách thức:

- Tính mơ hồ của ngôn ngữ.
- Hàm ý, ẩn ý.
- Ngôn ngữ gắn liền với tri thức.
- Sự phức tạp trong ngữ pháp.
- Giao tiếp thường là đa phương thức (multimodality).
- Thiếu ngữ liệu gán nhãn.

CÁC CƠ SỞ

NGÔN NGỮ HỌC

NGÔN NGỮ HỌC

Khái niệm về ngôn ngữ

Là **một hệ thống những đơn vị vật chất** và **những quy tắc hoạt động của chúng**, dùng làm công cụ giao tiếp của con người, được **phản ánh trong ý thức cộng đồng và trừu tượng hóa** khỏi bất kỳ một tư tưởng, cảm xúc và ước muốn cụ thể nào.

NGÔN NGỮ HỌC

Bản chất của ngôn ngữ

1. Hiện tượng xã hội đặc biệt.
2. Phương tiện giao tiếp quan trọng nhất của con người.
3. Hiện tượng trực tiếp của tư tưởng.
4. Phương tiện của tư duy.
5. *Hệ thống tín hiệu gồm có 2 mặt: mặt biểu hiện vật chất (âm, chữ) và mặt được biểu hiện (ý nghĩa).*

NGÔN NGỮ HỌC

Tính hệ thống của ngôn ngữ

- Các cấp độ trong ngôn ngữ

1. Âm vị (phoneme): **đơn vị âm thanh nhỏ nhất** để cấu tạo và khu biệt về mặt biểu hiện vật chất (âm thanh) của các đơn vị khác nhau. Ví dụ: b - i - g (big)
2. Hình vị (morpheme): **đơn vị nhỏ nhất mang nghĩa** (ngữ pháp hay từ vựng) được cấu tạo bởi các âm vị. Ví dụ: read-ing (reading)

NGÔN NGỮ HỌC

Tính hệ thống của ngôn ngữ

- Các cấp độ trong ngôn ngữ (tt)

3. Từ (word): đơn vị mang nghĩa độc lập, được cấu tạo bởi các hình vị, có chức năng định danh.
4. Ngữ (phrase): gồm hai hay nhiều từ có quan hệ ngữ pháp hay ngữ nghĩa với nhau.
5. Câu (sentence): các từ, ngữ có quan hệ ngữ pháp hay ngữ nghĩa với nhau và có chức năng cơ bản là thông báo

NGÔN NGỮ HỌC

Tính hệ thống của ngôn ngữ

- Các cấp độ trong ngôn ngữ (tt)

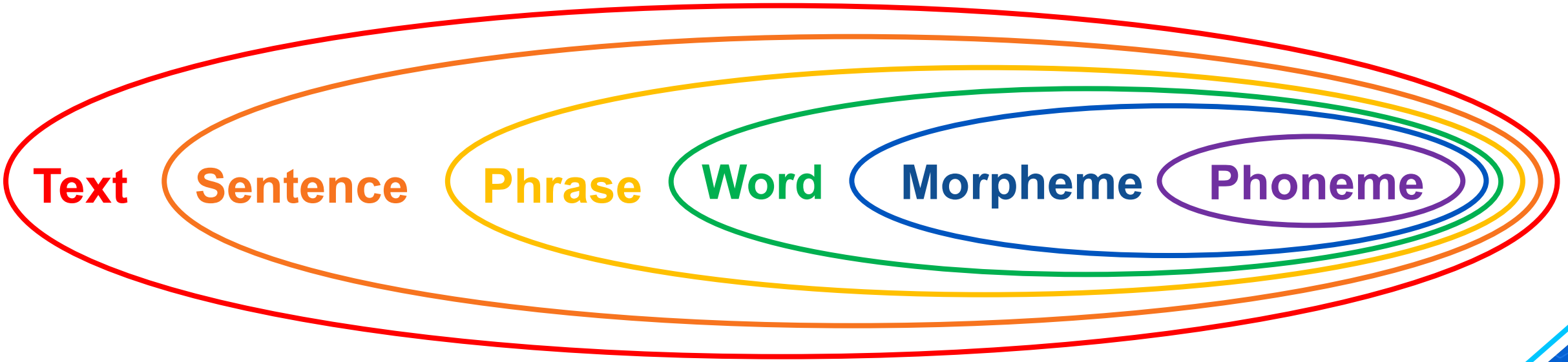
6. Văn bản (text): hệ thống các câu được liên kết với nhau về mặt hình thức, ngữ pháp, ngữ nghĩa và ngữ dụng.

NGÔN NGỮ HỌC

Tính hệ thống của ngôn ngữ

- Các quan hệ trong ngôn ngữ

1. Quan hệ cấp bậc (hierarchical relation): đơn vị cấp bậc cao hơn bao giờ cũng bao hàm đơn vị cấp bậc thấp hơn.



NGÔN NGỮ HỌC

Tính hệ thống của ngôn ngữ

- Các quan hệ trong ngôn ngữ (tt)

2. Quan hệ ngữ đoạn (syntagmatical relation): Nối kết các đơn vị ngôn ngữ thành chuỗi khi ngôn ngữ đi vào hoạt động (tính hình tuyến của ngôn ngữ). Các đơn vị ngôn ngữ phải nối tiếp nhau để tạo thành những sự kết hợp gọi là ngữ đoạn..

NGÔN NGỮ HỌC

Tính hệ thống của ngôn ngữ

- Các quan hệ trong ngôn ngữ (tt)

3. Quan hệ liên tưởng(associative relation): các yếu tố tương tự theo khía cạnh nào đó có thể thay thế nhau.

Ví dụ:

He saw a ***book*** ← {picture, pen, man, …}

NGÔN NGỮ HỌC

Tính hệ thống của ngôn ngữ

- Các phương diện trong ngôn ngữ

1. Hình thái (morphology): quan hệ giữa đơn vị ngôn ngữ với hình thức cấu tạo của nó
2. Ngữ pháp (syntax): quan hệ giữa đơn vị ngôn ngữ này với các đơn vị ngôn ngữ khác cùng xuất hiện với nó.
3. Ngữ nghĩa (semantics): quan hệ giữa đơn vị ngôn ngữ với nội dung (mặt ý nghĩa) của đơn vị đó

NGÔN NGỮ HỌC

Tính hệ thống của ngôn ngữ

- Các phương diện trong ngôn ngữ (tt)

4. Ngữ dụng (pragmatics): quan hệ giữa đơn vị ngôn ngữ với mục đích sử dụng của đơn vị đó.

NGÔN NGỮ HỌC

Phân loại ngôn ngữ

- Phân loại theo loại hình ngôn ngữ

1. Ngôn ngữ hòa kết (flexional): Đức, Latin, Anh, Pháp, …
2. Ngôn ngữ chắp dính (agglutinate) có hiện tượng nối tiếp thêm một hay nhiều phụ tố vào căn tố trong đó mỗi phụ tố chỉ mang một ý nghĩa ngữ pháp nhất định. (Nhật Bản, Triều Tiên, …)
3. Ngôn ngữ đơn lập (isolate): ngôn ngữ phi hình thái, không biến hình, đơn tiết, phân tiết. (Việt, Hán, …)

NGÔN NGỮ HỌC

Phân loại ngôn ngữ

- Phân loại theo trật tự từ

1. SVO: Anh, Việt, … chiếm 32.4 – 41.8%
2. SOV: Nhật, … chiếm 41 – 51.8%
3. VSO: chiếm 2 – 3%
4. VOS: chiếm 18%
5. OSV: chiếm khoảng 1%
6. OVS: chiếm khoảng 1%

CÁC CƠ SỞ

KHOA HỌC MÁY TÍNH

KHOA HỌC MÁY TÍNH

- Cấu trúc dữ liệu: chuỗi, cây, đồ thị, stack
- Thuật toán: tìm kiếm, phân tích cú pháp, tìm cây khung, tìm đường đi.
- Học máy xác suất (statistical learning)
- Học máy với mạng nơron (neural learning)

MỘT SỐ ỨNG DỤNG

CÁC BÀI TOÁN CƠ BẢN

CÁC BÀI TOÁN CƠ BẢN

1. Phân tách văn bản (tokenization)
2. **Gán nhãn từ loại (Part of Speech tagging – POS tagging)**
3. Nhận dạng thực thể (Named Entity Recognition – NER)
4. Phân tích hình thái (morphological analysis)
5. **Phân tích cú pháp (parsing)**
6. **Mô hình ngôn ngữ (language modeling)**
7. Phân tích ngữ nghĩa (semantic analysis)

MỘT SỐ ỨNG DỤNG

CHƯƠNG TRÌNH ỨNG DỤNG

CHƯƠNG TRÌNH ỨNG DỤNG

1. Sửa lỗi chính tả và ngữ pháp (spelling and grammar correction)
2. Hỏi-đáp tự động (question answering – QA)
3. Dịch máy (machine translation – MT)
4. Tóm tắt văn bản (text summarization)
5. Phân tích ý kiến (sentiment analysis – SA)
6. Chatbot.
7. …