Natural Language Processing

- Nguyễn Trung Hiếu –

Đại học Sư phạm Kỹ thuật Thành phố Hồ Chí Minh

Khoa CNTT

Email: hieuanthonydisward@gmail.com

Bài 1:Giới thiệu về NLP

1. Các cấp độ của phân tích:

Morphology: Cấu trúc bên trong của từ: ví dụ: unbreakable

- "un-" (tiền tố, nghĩa là "không")
- "break" (gốc từ, nghĩa là "phá vỡ")
- "-able" (hậu tố, nghĩa là "có thể")

Ứng dụng: Hiểu thêm nghĩa của từ(số ít, số nhiều, loại từ)

Nhưng không phải trường hợp nào cũng vậy

Ví dụ: arm: cánh tay, army: Quân đội

Syntax : Nghiên cứu cấu trúc ngữ pháp của câu, cách các từ được sắp xếp để tạo thành cụm từ và câu có nghĩa.

Ví dụ: Câu "The cat sat on the mat" có thể được phân tích cú pháp để xác định chủ ngữ ("the cat"), động từ ("sat"), và trạng ngữ ("on the mat").

Semantics: Nghiên cứu nghĩa của từ, cụm từ, và câu.

Dicourse: Nghiên cứu cách các câu được liên kết với nhau để tạo thành một đoạn văn hoặc bài văn mạch lạc.

Pragmatic: Nghiên cứu cách ngôn ngữ được sử dụng trong ngữ cảnh thực tế, bao gồm cả ý định của người nói và ảnh hưởng của ngữ cảnh đến nghĩa của lời nói

World Knowledge: Nghiên cứu cách ngôn ngữ được sử dụng trong ngữ cảnh thực tế, bao gồm cả ý định của người nói và ảnh hưởng của ngữ cảnh đến nghĩa của lời nói.

2. Word Segmentation

Có n cách để phân tách các từ trong một câu. Nhưng chỉ có một trong số chúng là đúng.

Lấy một chuỗi dài nhất từ một vị trí nào đó, chuỗi đó sẽ ở trong tự điển

Nhưng gây ra vấn đề lặp:

• Học sinh | học sinh | học.

• Học sinh | học | sinh học.

Kết nối các khả năng với nhau để tìm ra khả năng tốt nhất.

3. Gắn thẻ từ loại

The boy threw a ball to the brown dog.

- The/DT boy/NN threw/VBD a/DT ball/NN to/IN

the/DT brown/JJ dog/NN./.

- DT – determiner từ chỉ định
- NN – noun, danh từ, số ít hoặc số nhiều
- VBD – verb, past tense động từ, quá khứ
- IN – preposition giới từ
- JJ – adjective tính từ
- . – dấu chấm câu

Con ngựa đá con ngựa đá.

- Con ngựa/DT đá/ĐgT con ngựa/DT đá/DT.

Ông già đi nhanh quá

Ông/ĐaT già/TT đi/Phó_từ nhanh/TT

quá/trạng_từ.(ông già đi nhanh quá)

Ông già/DT đi/ĐgT nhanh/TT quá/trạng_từ. (Ông già/ đi nhanh quá)

Time // flies like an arrow.

Time flies // like an arrow.

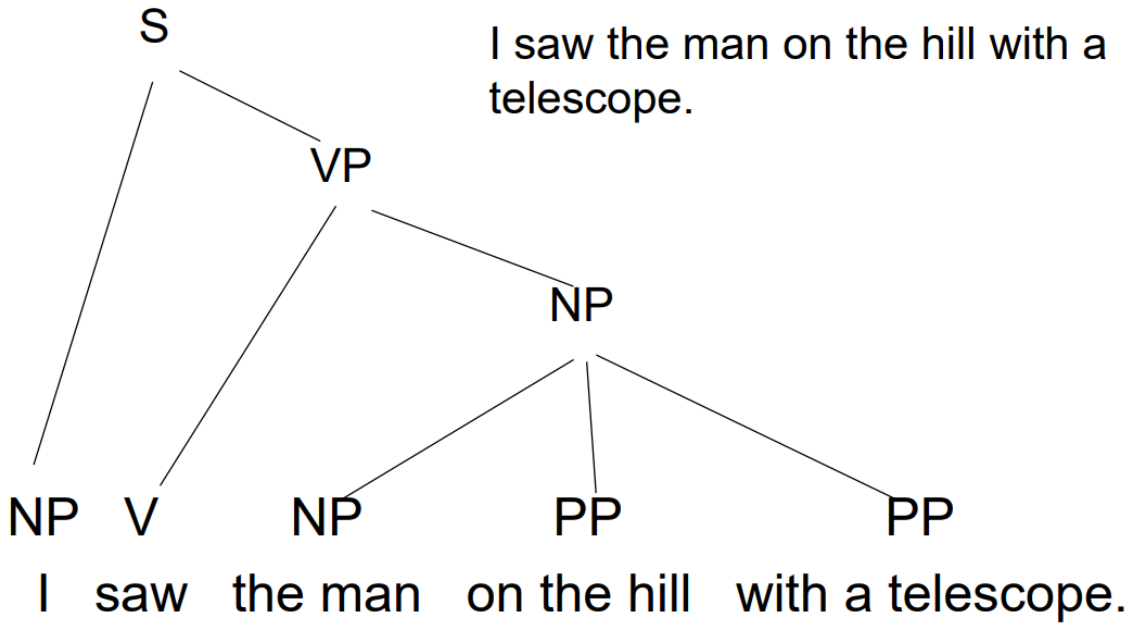
4. Cấu trúc mơ hồ

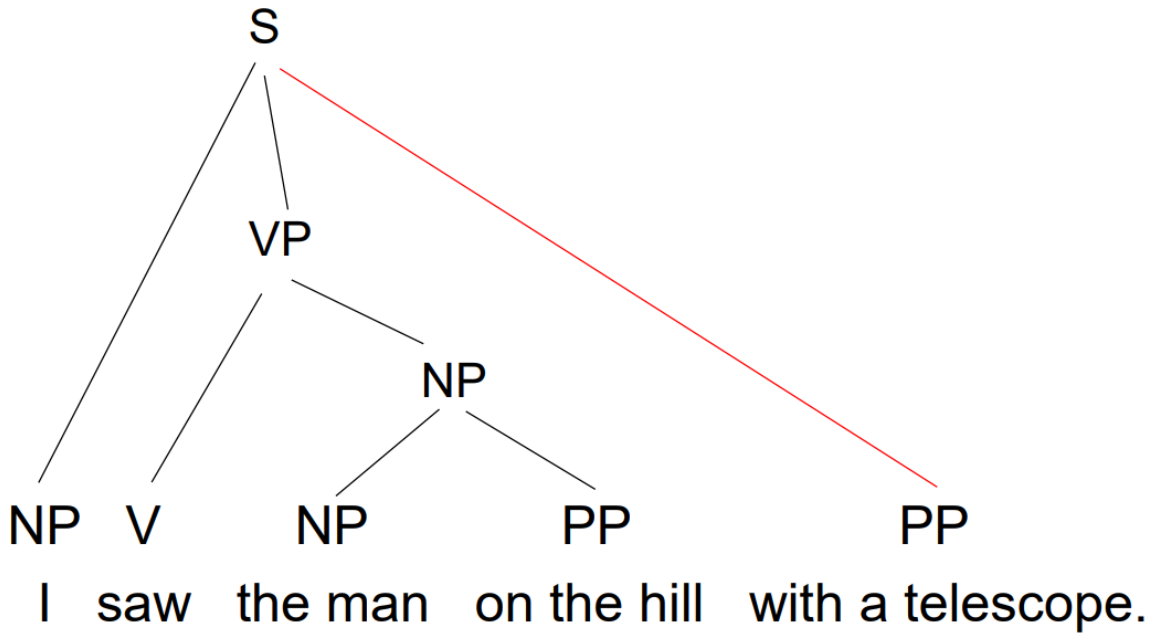
Xảy ra khi một câu có thể được phân tích cú pháp theo nhiều cách khác nhau, dẫn đến các hiểu nghĩa khác nhau. Sự mơ hồ này thường xuất hiện do cách các cụm từ hoặc mệnh đề được liên kết (attached) với nhau trong cấu trúc cú pháp của câu.

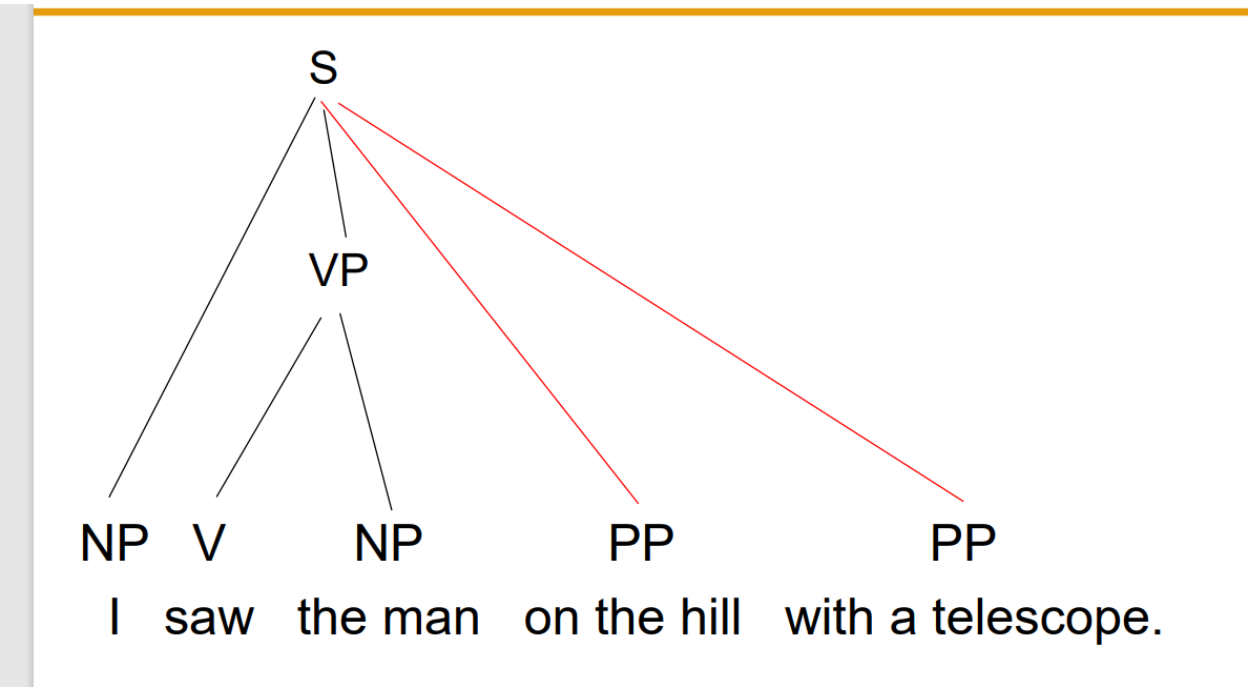
Chú thích:

- S: Câu
- VP: Cụm động từ
- NP: Cụm danh từ
- PP: Cụm giới từ

I saw the man on the hill with a telescope.







Ở đây ta có thể thấy 2 cụm giới từ(PP): On the hill và with a telescope

Giới từ là bổ nghĩa cho danh từ, nên câu này có thể hiểu theo nghĩa:

1. Tôi thấy người đàn ông trên đồi với kính viễn vọng(người đàn ông có kính viễn vọng)
2. Tôi dùng kính viễn vọng để nhìn người đàn ông(Tôi thấy người đàn ông qua kính viễn vọng)

- Cú pháp (syntax) chỉ cho chúng ta biết về cấu trúc ngữ pháp của câu, tức là cách các từ được sắp xếp và liên hệ với nhau để tạo thành câu đúng ngữ pháp. Tuy nhiên, cú pháp không đảm bảo rằng câu đó có nghĩa, hợp lý, hoặc dễ hiểu.

Ví dụ:

Colorless green ideas sleep furiously:

Những ý tưởng xanh không màu(chủ ngữ) ngủ(động từ) một cách dữ dội(trạng ngữ) -> đúng cú pháp nhưng lại không hiểu về nghĩa. Sao xanh lại không màu được?

Phải kết hợp cả cú pháp lẫn ngữ nghĩa, các hệ thống đều sử dụng các phương pháp:

- Phân tích cú pháp (Parsing): Để xác định cấu trúc ngữ pháp.
- Biểu diễn ngữ nghĩa (Semantic representation): Sử dụng các mô hình như word embeddings (word2vec, GloVe), sentence embeddings (BERT, Sentence-BERT) để biểu diễn ý nghĩa của từ và câu.
- Kiến thức thế giới (World knowledge): Sử dụng cơ sở tri thức (knowledge base) hoặc các mô hình ngôn ngữ lớn được huấn luyện trên lượng lớn dữ liệu để hiểu mối quan hệ giữa các khái niệm.

5. Mơ hồ trong từ vựng

Ví dụ như:

• I walked to the bank ...

of the river/to get money.

• I work for John Hancock ...

and he is a good boss/which is a good company

<u>Cách xử lý:</u>

Word Sense Disambiguation (WSD - Phân biệt nghĩa của từ): Đây là nhiệm vụ xác định nghĩa chính xác của một từ trong một ngữ cảnh cụ thể. Các phương pháp WSD sử dụng nhiều kỹ thuật khác nhau, bao gồm:

- **Dựa trên từ điển và cơ sở tri thức (Dictionary- and Knowledge-based):** Sử dụng thông tin từ WordNet hoặc các cơ sở tri thức khác để xác định các nghĩa có thể của từ và lựa chọn nghĩa phù hợp nhất dựa trên ngữ cảnh.

- **Dựa trên thống kê (Statistical-based):** Sử dụng các mô hình thống kê được huấn luyện trên dữ liệu lớn để dự đoán nghĩa của từ dựa trên các từ xung quanh.
- **Dựa trên học sâu (Deep learning-based):** Sử dụng mạng nơ-ron để học các biểu diễn ngữ cảnh của từ và phân biệt nghĩa của chúng.

Ngữ cảnh (Context): Sử dụng thông tin từ các từ và câu xung quanh để xác định nghĩa phù hợp nhất.

Kiến thức thế giới (World knowledge): Sử dụng kiến thức về thế giới thực để hiểu mối quan hệ giữa các khái niệm và giải quyết sự mơ hồ.

6. Đồng tham chiếu

Coreference là mối quan hệ giữa các biểu thức ngôn ngữ (thường là cụm danh từ) cùng tham chiếu đến một thực thể duy nhất trong thế giới thực. Nói cách khác, các biểu thức khác nhau nhưng chỉ cùng một người, vật, hoặc khái niệm.

President John F. Kennedy was assassinated.

The president was shot yesterday.

Relatives said that John was a good father.

JFK was the youngest president in history.

His family will bury him tomorrow.

Tất cả các từ được bôi đỏ đều chỉ về một người đó là John F. Kenedy

<u>Ứng dụng của đồng tham chiếu:</u>

- **Tóm tắt văn bản (Text summarization):** Để tóm tắt chính xác, hệ thống cần biết các biểu thức nào đề cập đến cùng một thực thể.
- **Trả lời câu hỏi (Question answering):** Để trả lời các câu hỏi liên quan đến một thực thể được đề cập trong văn bản, hệ thống cần xác định tất cả các lần xuất hiện của thực thể đó.

- **Dịch máy (Machine translation):** Để dịch chính xác, đặc biệt là các đại từ, hệ thống cần biết chúng tham chiếu đến ai.
- **Khai thác thông tin (Information extraction):** Để trích xuất thông tin về các thực thể, hệ thống cần xác định tất cả các lần đề cập đến chúng.

Các loại của đồng tham chiếu:

- **Anaphora:** Một biểu thức (thường là đại từ) tham chiếu đến một biểu thức được đề cập trước đó trong văn bản (ví dụ: "John went to the store. He bought milk.").
- **Cataphora**: Một biểu thức tham chiếu đến một biểu thức được đề cập sau đó trong văn bản (ví dụ: "When he arrived, John went to the store."). (Ít phổ biến hơn Anaphora)
- **Pronominal coreference:** Đồng tham chiếu sử dụng đại từ (ví dụ: he, she, it, they).
- **Nominal coreference:** Đồng tham chiếu sử dụng cụm danh từ (ví dụ: "the president", "Mr. Kennedy").

Cách giải quyết Coreference Resolution:

- Dựa trên quy tắc (Rule-based): Sử dụng các quy tắc ngữ pháp và ngữ nghĩa để xác định đồng tham chiếu.
- Dựa trên thống kê (Statistical-based): Sử dụng các mô hình thống kê được huấn luyện trên dữ liệu lớn.
- Dựa trên học sâu (Deep learning-based): Sử dụng mạng nơ-ron để học các biểu diễn ngữ cảnh và giải quyết đồng tham chiếu.

7. Ngữ dụng học(Pragmatics)

Ngữ dụng học nghiên cứu cách ngôn ngữ được sử dụng trong ngữ cảnh thực tế và cách người nghe diễn giải ý nghĩa của người nói. Nó vượt ra ngoài nghĩa đen của từ và câu để xem xét ý định của người nói, tác động của lời nói đến người nghe, và các quy tắc chi phối giao tiếp.

Ứng dụng:

- **Hiểu được ý định của người nói:** Người nói muốn gì khi nói câu đó? Họ đang yêu cầu, đề nghị, hỏi thông tin, hay bày tỏ cảm xúc?
- **Phản ứng của người nghe:** Người nghe hiểu lời nói đó như thế nào? Họ có đồng ý, phản đối, hay thắc mắc?
- **Ngữ cảnh giao tiếp:** Cuộc trò chuyện diễn ra ở đâu, giữa những ai, và mục đích của cuộc trò chuyện là gì?

Ví dụ, nếu ai đó nói "Trời nóng quá!", tùy vào ngữ cảnh, nó có thể mang nhiều ý nghĩa:

Than phiền: Người nói đang cảm thấy khó chịu vì thời tiết nóng.

Đề nghị mở điều hòa: Người nói muốn người nghe mở điều hòa.

Gợi ý đi bơi: Người nói muốn rủ người nghe đi bơi.

Các quy tắc hội thoại:

Đây là những quy tắc ngầm định chi phối cách chúng ta giao tiếp với nhau. Paul Grice đã đề xuất "Maxims of Conversation" (Phương châm Hội thoại) bao gồm:

- **Maxim of Quantity (Phương châm về Lượng):** Đưa ra thông tin vừa đủ, không quá nhiều cũng không quá ít.
- **Maxim of Quality (Phương châm về Chất):** Nói sự thật, không nói điều mình tin là sai hoặc thiếu bằng chứng.
- **Maxim of Relation (Phương châm về Liên quan):** Nói những điều liên quan đến chủ đề đang nói.
- **Maxim of Manner (Phương châm về Cách thức):** Nói rõ ràng, mạch lạc, tránh mơ hồ và khó hiểu.

Các hành động ngôn ngữ:

- **Lời khẳng định (Assertives):** Khẳng định một điều gì đó là đúng (ví dụ: "Trời đang mưa.").
- **Lời chỉ thị (Directives):** Yêu cầu người nghe làm một điều gì đó (ví dụ: "Đóng cửa lại!").
- **Lời cam kết (Commissives):** Cam kết thực hiện một hành động trong tương lai (ví dụ: "Tôi sẽ đến đúng giờ.").
- **Lời biểu lộ (Expressives):** Bày tỏ cảm xúc hoặc thái độ (ví dụ: "Chúc mừng sinh nhật!").
- **Lời tuyên bố (Declarations):** Thay đổi trạng thái của thế giới thông qua lời nói (ví dụ: "Tôi tuyên bố cuộc họp kết thúc.").

Ứng dụng trong Chatbox:

- **Hiểu được ý định của người nói:** Điều này quan trọng cho các ứng dụng như trợ lý ảo, chatbot.
- **Xử lý hội thoại một cách tự nhiên hơn**: Các hệ thống cần tuân theo các quy tắc hội thoại để giao tiếp hiệu quả.
- **Phân tích tình cảm và thái độ:** Ngữ dụng học giúp hiểu được cảm xúc và thái độ được thể hiện qua lời nói.

8. Hiểu biết kiến thức thế giới(World Knowledge)

Một câu nói sẽ chứa đựng nhiều thông tin ngầm định mà chúng ta, với kiến thức về thế giới, có thể dễ dàng suy luận ra. Tuy nhiên, một hệ thống NLP chỉ dựa vào cú pháp và ngữ nghĩa đơn thuần sẽ gặp khó khăn.

Ví dụ:

Mai went to the diner. She ordered a steak. She left a tip and went home.

- **What did Mai eat for dinner?**

Trong văn bản đề cập Mai gọi bít tết. Con người dễ dàng hiểu là cô ấy đã dùng món bít tết.

Tuy nhiên việc Mai ăn bít tết không được nêu ra -> Con người phải suy luận ra

• Who brought Mai her food?

Không có đề cập cho thấy người mang món ăn ra cho Mai, nhưng với hiểu biết về thế giới chúng ta có thể suy luận được rằng bồi bàn hoặc nhân viên đã mang ra.

• Who cooked the steak?

Không có đề cập cho thấy người nấu món bít tết, nhưng với hiểu biết về thế giới chúng ta có thể suy luận được đầu bếp nấu món đó

• Did Mai pay her bill?

Văn bản nói "She left a tip" (Cô ấy để lại tiền boa).

Kiến thức thế giới cho chúng ta biết rằng tiền boa (tip) thường được để lại sau khi đã thanh toán hóa đơn (bill) cho dịch vụ.

Do đó, chúng ta có thể suy luận rằng Mai đã trả hóa đơn trước khi để lại tiền boa.

Các hệ thống NLP cần trang bị các công cụ để hiểu và suy luận từ ngôn ngữ:

- **Suy luận (Inference):** Rút ra kết luận dựa trên thông tin được cung cấp và kiến thức đã biết.
- **Giải quyết sự mơ hồ (Ambiguity resolution):** Xác định nghĩa chính xác của từ hoặc cụm từ dựa trên ngữ cảnh và kiến thức thế giới.
- **Hiểu được ý định của người viết/nói (Understanding author's/speaker's intention):** Hiểu được mục đích thực sự của thông điệp.

9. Kiến thức cơ bản về ngôn ngữ

<u>Quy tắc 1:</u> Từ phải xuất hiện theo một trật tự cụ thể

Ví dụ:

Chó kem ăn.(Sai)

Chó ăn kem.(Đúng)

Cấu trúc phải đúng: S(chủ ngữ) + V(Động từ) + O(Tân ngữ)

<u>Quy tắc 2:</u> Các bộ phận cấu thành câu

Có đủ chủ ngữ vị ngữ

<u>Quy tắc 3:</u> Ai đã làm gì với ai/cái gì?

Đây là một cách diễn đạt rất tốt để tóm tắt ý nghĩa của câu và mối quan hệ giữa các thành phần. Trong câu "Chó ăn kem":

- chủ thể (chó): là người/vật thực hiện hành động (who).
- hành động (ăn): là việc được thực hiện (what).
- đối tượng (kem): là người/vật chịu tác động của hành động (to whom/what).

10. Kiến thức ẩn

Trong NLP (Xử lý Ngôn ngữ Tự nhiên), "Hidden Knowledge" (Kiến thức ẩn) đề cập đến những hiểu biết ngầm về ngôn ngữ mà con người có được một cách tự nhiên, thường là vô thức, và được sử dụng để hiểu và xử lý ngôn ngữ. Nó khác với kiến thức tường minh (explicit knowledge) được dạy hoặc học một cách rõ ràng.

Các kiểu kiến thức ẩn:

- **Kiến thức về cú pháp (Syntactic knowledge):** Hiểu biết về cấu trúc câu, trật tự từ, và các quy tắc ngữ pháp. Ví dụ, chúng ta biết rằng câu "Chó ăn kem" đúng ngữ pháp, trong khi "Kem ăn chó" thì không.

- **Kiến thức về ngữ nghĩa (Semantic knowledge):** Hiểu biết về nghĩa của từ và cách chúng kết hợp với nhau để tạo thành nghĩa của câu. Ví dụ, chúng ta biết rằng "ăn" liên quan đến hành động tiêu thụ thức ăn, và "kem" là một loại thức ăn.
- **Kiến thức về ngữ cảnh (Contextual knowledge):** Hiểu biết về ngữ cảnh sử dụng ngôn ngữ, bao gồm tình huống giao tiếp, người nói, người nghe, và mục đích của giao tiếp. Ví dụ, câu "Trời nóng quá!" có thể mang nhiều ý nghĩa khác nhau tùy thuộc vào ngữ cảnh.
- **Kiến thức về thế giới (World knowledge):** Hiểu biết về thế giới thực, bao gồm các sự kiện, con người, địa điểm, và mối quan hệ giữa chúng. Ví dụ, chúng ta biết rằng người phục vụ mang thức ăn cho khách trong nhà hàng.
- **Kiến thức về diễn ngôn (Discourse knowledge)**: Hiểu biết về cách các câu liên kết với nhau để tạo thành một đoạn văn mạch lạc. Ví dụ, chúng ta có thể hiểu mối quan hệ đồng tham chiếu giữa "John" và "ông ấy" trong một đoạn văn.

Ví dụ:

- **"I want to solve the problem" vs. "I wanna solve the problem"**: Kiến thức ẩn về cách sử dụng dạng rút gọn "wanna" trong tiếng Anh, cho biết nó được sử dụng trong ngữ cảnh thân mật hơn.
- **"I understand these students" vs. "These students I understand"**: Kiến thức ẩn về trật tự từ và quy tắc hòa hợp giữa chủ ngữ và động từ trong tiếng Anh.
- **"I want these students to solve the problem" vs. "These students I want [x] to solve the problem"**: Kiến thức ẩn về cấu trúc câu phức và khả năng suy luận ra thành phần bị thiếu dựa trên ngữ cảnh.

Vai trò trong NLP:

- **Phân tích cú pháp (Parsing):** Xác định cấu trúc ngữ pháp của câu.
- **Phân tích ngữ nghĩa (Semantic analysis):** Hiểu ý nghĩa của câu.
- **Giải quyết sự mơ hồ (Ambiguity resolution):** Xác định nghĩa chính xác của từ hoặc cụm từ trong ngữ cảnh.
- **Suy luận (Inference):** Rút ra kết luận dựa trên thông tin được cung cấp.
- **Hiểu diễn ngôn (Discourse understanding):** Hiểu mối quan hệ giữa các câu trong một đoạn văn.

11. LSAT / (former) GRE Analytic Section Questions

Này là bài tập

Các điều kiện:

- Sáu tác phẩm điêu khắc: C, D, E, F, G, H
- Ba phòng trưng bày: 1, 2, 3
- C và E không được ở cùng phòng.
- D và G phải ở cùng phòng.
- Nếu E và F ở cùng phòng, không tác phẩm nào khác được ở trong phòng đó.
- Mỗi phòng có ít nhất một tác phẩm, và không quá ba tác phẩm trong một phòng.

Tình huống cụ thể:

- D ở phòng 3
- E và F ở phòng 1

Phân tích và loại trừ:

1. **D ở phòng 3:** Vì D và G phải ở cùng phòng, nên G cũng ở phòng 3. Vậy phòng 3 có D và G.
2. **E và F ở phòng 1:** Theo điều kiện, nếu E và F ở cùng phòng, thì không tác phẩm nào khác được ở trong phòng đó. Vậy phòng 1 chỉ có E và F.
3. **Còn lại:** Chúng ta còn tác phẩm C và H, và phòng 2 vẫn trống. Vì mỗi phòng phải có ít nhất một tác phẩm, nên C và H phải ở phòng 2.

Kiểm tra các lựa chọn:

- **A. Sculpture C is exhibited in room 1 (Tác phẩm C được trưng bày ở phòng 1):** Sai. C ở phòng 2.
- **B. Sculpture H is exhibited in room 1 (Tác phẩm H được trưng bày ở phòng 1):** Sai. H ở phòng 2.
- **C. Sculpture G is exhibited in room 2 (Tác phẩm G được trưng bày ở phòng 2):** Sai. G ở phòng 3.
- **D. Sculptures C and H are exhibited in the same room (Tác phẩm C và H được trưng bày trong cùng một phòng):** Đúng. Cả hai đều ở phòng 2.
- **E. Sculptures G and F are exhibited in the same room (Tác phẩm G và F được trưng bày trong cùng một phòng):** Sai. G ở phòng 3, F ở phòng 1.

Kết luận:

Đáp án đúng là **D**.

12. Giải quyết tham chiếu

Xác định các biểu thức ngôn ngữ (từ, cụm từ) trong một đoạn hội thoại hoặc văn bản tham chiếu đến cùng một đối tượng hoặc khái niệm nào.

Nói một cách đơn giản, nó giúp máy tính hiểu được "cái gì" hoặc "ai" mà các đại từ, cụm danh từ, hoặc các biểu thức khác đang đề cập đến.

Các cách giải quyết tham chiếu:

- **Domain knowledge (Kiến thức miền):** Kiến thức về một lĩnh vực cụ thể. Ví dụ, trong đoạn hội thoại về rạp chiếu phim, kiến thức về cách hoạt động của rạp chiếu phim (có lịch chiếu, giá vé, v.v.) là kiến thức miền.
- **Discourse knowledge (Kiến thức diễn ngôn):** Kiến thức về cấu trúc và sự mạch lạc của một đoạn hội thoại hoặc văn bản. Ví dụ, chúng ta biết rằng câu trả lời thường liên quan đến câu hỏi trước đó.
- **World knowledge (Kiến thức thế giới):** Kiến thức chung về thế giới, bao gồm các sự kiện, con người, địa điểm, và mối quan hệ giữa chúng. Ví dụ, chúng ta biết rằng người lớn thường trả vé xem phim khác với trẻ em.

Ví dụ:

Đoạn hội thoại giữa người dùng (U) và hệ thống (S) về bộ phim "A Bug's Life" minh họa cách các nguồn kiến thức này được sử dụng để giải quyết tham chiếu:

- **U: Where is A Bug's Life playing in Mountain View? (Phim A Bug's Life đang chiếu ở đâu tại Mountain View?)**
    - Câu hỏi này giới thiệu đối tượng chính: bộ phim "A Bug's Life" và địa điểm "Mountain View".
- **S: A Bug's Life is playing at the Summit theater. (Phim A Bug's Life đang chiếu tại rạp Summit.)**
    - Hệ thống trả lời bằng cách cung cấp địa điểm cụ thể: "the Summit theater".
- **U: When is *it* playing *there*? (Khi nào *nó* chiếu *ở đó*?)**

    - **"it"** tham chiếu đến "A Bug's Life" (kiến thức diễn ngôn: câu hỏi liên quan đến câu trước).
    - **"there"** tham chiếu đến "the Summit theater" (kiến thức diễn ngôn: câu hỏi liên quan đến câu trước).
- **S: It's playing at 2pm, 5pm, and 8pm. (Nó chiếu vào lúc 2 giờ chiều, 5 giờ chiều và 8 giờ tối.)**
    - **"It"** tiếp tục tham chiếu đến "A Bug's Life" (kiến thức diễn ngôn).
- **U: I'd like 1 adult and 2 children for the first show. How much would *that* cost? (Tôi muốn 1 vé người lớn và 2 vé trẻ em cho suất chiếu đầu tiên. *Đó* sẽ tốn bao nhiêu?)**
    - **"that"** tham chiếu đến "1 adult and 2 children for the first show" (kiến thức diễn ngôn: câu hỏi liên quan đến yêu cầu đặt vé).
    - Để trả lời câu hỏi "How much would that cost?", hệ thống cần kiến thức miền (giá vé người lớn và trẻ em) và kiến thức thế giới (vé trẻ em thường rẻ hơn vé người lớn).

13. Hình thái học

Ví dụ:
Singing: Sing + ing: Hiện tại tiếp diễn

Tuy nhiên

Duckling-> Duckl + ing

Duckl không có nghĩa

Vì vậy không nên học thuộc và áp dụng máy móc. Nên biết rằng Duckl không có nghĩa.

Phải biết các quy tắc:
Book + s -> Books

Box + s -> boxes

14. Đặc điểm của NLP

1. Ambiguous at all levels (Mơ hồ ở mọi cấp độ)

Tính mơ hồ là một đặc điểm cố hữu của ngôn ngữ tự nhiên. Một câu hoặc một từ có thể được hiểu theo nhiều cách khác nhau tùy thuộc vào ngữ cảnh. Sự mơ hồ này tồn tại ở nhiều cấp độ:

- **Mơ hồ về từ vựng (Lexical Ambiguity):** Một từ có thể có nhiều nghĩa.
    - Ví dụ: Từ "bank" trong tiếng Anh có thể có nghĩa là "ngân hàng" (nơi giao dịch tiền bạc) hoặc "bờ sông". Trong tiếng Việt, từ "cờ" có thể là "lá cờ" (biểu tượng) hoặc "cờ tướng/cờ vua" (môn thể thao).
- **Mơ hồ về cú pháp (Syntactic Ambiguity):** Cấu trúc câu có thể cho phép nhiều cách hiểu.
    - Ví dụ: "Tôi thấy người đàn ông với cái kính viễn vọng." Câu này có thể hiểu là "Tôi dùng kính viễn vọng để nhìn thấy người đàn ông" hoặc "Tôi nhìn thấy người đàn ông đang cầm/có cái kính viễn vọng". Trong tiếng Việt, "Con chó của người hàng xóm cắn người đưa thư." có thể hiểu là "Con chó (thuộc sở hữu của người hàng xóm) cắn người đưa thư" hoặc "Con chó cắn người đưa thư (là người hàng xóm)".
- **Mơ hồ về ngữ nghĩa (Semantic Ambiguity):** Ý nghĩa của cả câu có thể không rõ ràng do sự mơ hồ của các từ hoặc cụm từ trong câu.

- Ví dụ: "Đi thăm người thân có thể mệt mỏi." Có thể hiểu là "Việc đi thăm người thân khiến người ta mệt mỏi" hoặc "Những người thân đến thăm khiến người ta mệt mỏi".

2. Involve reasoning about the world (Liên quan đến suy luận về thế giới)

Để hiểu ngôn ngữ một cách đầy đủ, chúng ta thường cần suy luận về thế giới và ngữ cảnh mà ngôn ngữ được sử dụng. Điều này bao gồm:

- **Kiến thức thông thường (Common sense knowledge):** Hiểu biết về các khái niệm hàng ngày và mối quan hệ giữa chúng.
    - Ví dụ: Chúng ta biết rằng chim có thể bay, người ăn thức ăn, và nước thì ướt.
- **Kiến thức thế giới (World knowledge):** Hiểu biết về các sự kiện và thông tin cụ thể về thế giới.
    - Ví dụ: Chúng ta biết rằng Hà Nội là thủ đô của Việt Nam, hoặc mặt trời mọc ở hướng đông.
- **Kiến thức ngữ cảnh (Contextual knowledge):** Hiểu biết về tình huống cụ thể và thông tin nền liên quan đến cuộc trò chuyện.
    - Ví dụ: Trong một cuộc trò chuyện về bóng đá, chúng ta cần biết tên các đội bóng và kết quả trận đấu để hiểu được nội dung cuộc trò chuyện.

Ví dụ minh họa kết hợp cả hai đặc điểm:

"Cô ấy nhìn thấy con dơi trên cây bằng ống nhòm."

- **Mơ hồ về cú pháp:** "bằng ống nhòm" có thể bổ nghĩa cho "nhìn thấy" (cô ấy dùng ống nhòm để nhìn) hoặc "con dơi" (con dơi có ống nhòm).
- **Suy luận về thế giới:** Nếu ngữ cảnh là cô ấy đang đi thám hiểm trong rừng, khả năng cao là "bằng ống nhòm" bổ nghĩa cho "nhìn

thấy". Ngược lại, nếu ngữ cảnh là một câu chuyện viễn tưởng về một con dơi có khả năng đặc biệt, thì "bằng ống nhòm" có thể bổ nghĩa cho "con dơi".

(Ghi chú được tham khảo từ slide môn xử lý ngôn ngữ tự nhiên của HUST)

Nguồn Slide: https://github.com/hoanglong1712/Dai-Hoc-Bach-Khoa-Ha-Noi/tree/main/Natural%20Language%20Processing