Tách từ và gán nhãn từ loại là hai bước tiền xử lý rất quan trọng trong việc xây dựng dữ liệu để thực hiện một loạt các công việc của xứ lý ngôn ngữ tự nhiên như gán nhãn quan hệ cú pháp phụ thuộc, cú pháp thành phần,… Do các dữ liệu thô được khai thác chủ yếu ở trên các trang mạng nên việc gặp phải các trường hợp đặc biệt, nhập nhằng trong xử lý tách từ cũng tạp ra không ít trở ngại cho việc tiền xử lý. Cách tách từ và gán nhãn từ loại cũng thay đổi theo thời gian để đảm bảo tính phù hợp và đồng bộ hơn trong việc xây dựng dữ liệu. Các thay đổi trong tách từ và gán nhãn sẽ được chúng tôi tổng hợp một cách hết sức ngắn gọn dưới dây.

Về nhãn từ loại, các nhãn từ loại cũ sẽ được thay thế bằng các nhãn từ loại mới theo bảng sau.

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **POS\_old** | **POS\_2022** | **vnPOS** |
|  | N | N | danh từ chung (Noun) |
|  | Np | NNP | danh từ riêng (Proper noun) |
|  | Nc | NC | danh từ loại thể (Classifier noun)  [nhập chung Nc, Ns, Ne thành Nc] |
|  | Nu | NU | danh từ chỉ đơn vị đo lường, tiền tệ (Unit Noun) |
|  | Nux | NUX | tổ hợp danh từ chỉ đơn vị mở rộng (Extended unit noun) |
|  | M | NUM | số từ (Numeral) |
|  | Mx | NUMX | số từ mở rộng (Extended numeral) |
|  | L, Q | DET | lượng từ, từ định lượng, định từ (Determiner) |
|  | V | V | động từ (Verb) |
|  | V | AUX | trợ động từ, hệ từ LÀ (Auxiliary Verb) |
|  | A | ADJ | tính từ (Adjective) |
|  | P | PRO | đại từ (Pronoun) |
|  | R | ADV | phụ từ (Adverb) |
|  | E | PRE | giới từ (Preposition) |
|  | Cc | CC | liên từ phối hợp (Coordinating Conjunction) |
|  | C | SC | liên từ phụ thuộc (Subordinating Conjunction) |
|  | T, Aux | PRT | trợ từ (Particle) |
|  | I | I | cảm từ (Interjection) |
|  | G, X | ADJ | tổ hợp có thính chất như tính từ |
|  | G, X | ADV | tổ hợp có thính chất như phụ từ, bổ nghĩa cho động từ, tính từ |
|  | G, X | N | tổ hợp có thính chất như danh từ |
|  | G, X | PRT | tổ hợp có thính chất như trợ từ |
|  | G, X | SC | tổ hợp có thính chất như liên từ phụ thuộc |
|  | G, X | V | tổ hợp có thính chất như động từ |
|  | MW | MWE | ngữ cố định/thành ngữ (Multi-word expression) |
|  | D | D | tổ hợp từ biểu thị ý nghĩa tục tĩu (Dirty word) |
|  | X | X | tổ hợp không thể xác định từ loại (Undetermined group) |
|  | Z | Z | yếu tố cấu tạo từ (Bound morphemes) |
|  | y | y | Từ viết tắt (Abbreviation, Acronym) |
|  | b | b | Từ vay mượn (Borrowed words) |
|  | FW | FW | Từ nguyên dạng tiếng nước ngoài (Foreign words) |
|  | CH, SYM | PUNCT | Dấu câu (Punctuation) |
|  | CH | SYM | Ký hiệu (Symbol) |

Bên cạnh sự thay đổi về quy ước gán nhãn từ loại, còn có một số thay đổi như sau:

* Nhãn X không còn xuất hiện trong dữ liệu, do dữ liệu đều được gán nhãn từ loại phụ hợp với các vai trò ngữ pháp của từ và cụm từ. Như vậy nhãn X sẽ được gán nhãn bằng các nhãn từ loại có tính chất tương ứng.

VD: một\_khác (X) – một khác (ADJ)

* Giống như nhãn X, nhãn G cũng không còn xuất hiện mà được thay thế bằng các nhãn (ADJ, ADV, N, SC, PRT, V) theo đúng tính chất của tổ hợp đó. VD: nói\_tóm\_lại (G) – nói\_tóm\_lại (V)
* Nhãn MW được thay thế bằng nhãn MWE – ngữ cố định/thành ngữ. Mỗi đơn vị từ trong ngữ cố định và tổ hợp từ thì lại được gán từ loại bình thường, cả tổ hợp từ đó được đánh dấu với các nhãn con.

VD: hang\_cùng\_ngõ\_hẻm (MW) – (NP-MWE (N hang) (ADJ cùng) (N ngõ) (ADJ hẻm))

=> Tức là các đơn vị từ trong tổ hợp được gán nhãn từ loại riêng, và đến phần mình, tổ hợp từ được gán nhãn là NP – một danh ngữ.

* Nhãn Z – yếu tố cấu tạo từ không xuất hiện trong dữ liệu, do các yếu tố cấu tọa từ được ghép với các từ đi kèm để tạo thành một đơn vị từ.

VD: siêu\_người\_mẫu (N). Trong đó siêu (Z) có nghĩa là ‘cao vượt lên trên’

* Một thay đổi nữa được chúng tôi đánh giá là đơn giản nhưng rất quan trong đó là sự thay đổi về quy ước giữa dấu câu và ký hiệu. Với các dấu câu, trong quy ước gán nhãn cũ sẽ được gán là CH, SYM, trong cách gán nhãn mới được gán là PUNCT. Còn về ký hiệu, cách gán cũ là CH, cách gán mới là SYM. Điều này sẽ giúp khắc phục tình trạng nhập nhằng, không nhất quán khi gán nhãn từ loại cho các dấu câu và ký hiệu.

Về tách từ, cũng không có quá nhiều thay đổi so với cách làm trước năm 2022. Tuy nhiên, chúng ta vẫn phải kể đến một số thay đổi trong tách từ như sau:

* Thứ nhất, về cách gán nhãn danh từ riêng.
* Với tên người: cách gán nhãn tên người được giữ nguyên khi xử lý cụm tên người, coi mỗi thành tố của tên người là một từ và viết hoa các chữ cái đầu của các thành tố, các nhãn NNP và NNPy được gán nhãn cho các thành tố.

VD: Phạm/Np Văn/Np Đồng/Np – Phạm/NNP Văn/NNP Đồng/NNP

* Với tên người được riêng hóa từ danh từ chung thì trước và sau năm 2022 sẽ có sự khác biệt.

VD: La/Np Thị/Np Cẩm\_Chướng/Np – La/NNP Thị/NNP Cẩm/NNP Chướng/NNP (cẩm chướng là danh từ chung)

* Tên địa lý: Khi xử lý các cụm tên địa lý, cách gán nhãn trước và sau 2022 có sự khác biệt. Nếu như trước năm 2022 không tách các yếu tố trong tên địa lý thì sau năm 2022 có xu hướng xử lý như xử lý đối với cụm tên riêng chỉ người.

VD: TP/N Hồ\_Chí\_Minh/Np – TP/N Hồ/NNP Chí/NNP Minh/NNP

* Thứ hai, khi xử lý các ngữ cố định, thành ngữ, chúng ta phải tách các trong cụm ra và gán cho mỗi từ riêng lẻ một nhãn từ loại khác nhau.

VD: chuột\_chạy\_cùng\_sào – chuột/N chạy/V cùng/N sào/N

* Cuối cùng, chúng ta phải chú ý đến hiện tượng nhập nhằng trong việc xử lý các cụm có các yếu tố có thể hoặc không là yếu tố cấu tạo từ. Từ “phó” là một ví dụ tiêu biểu cho nhãn Z, khi tham gia vào cấu tạo từ trong “phó trưởng phòng”, “phó” không thể đứng riêng lẻ mà phải được gộp vào trong cụm, chính vì vậy cả cụm sẽ được xử lý là “phó\_trưởng\_phòng”. Nhìn chung là cách xử lý này khá thống nhất trước và sau 2022. Có một lưu ý khi xử lý những cụm đặc biệt khác. Ví dụ “trưởng phòng kế hoạch” từ việc xác định từ loại của “trưởng” rất cần được lưu ý. Từ “trưởng” không phải mang nhãn Z mà mang nhãn N, hoạt động độc lập và hoàn “phòng kế hoạch” có quan hệ phụ thuộc cho “trưởng”. Cụm sẽ được xử lý tách từ và gán nhãn như sau “trưởng/N phòng\_kế\_hoạch/N”