TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO CUỐI KÌ MÔN NHẬP MÔN**

**XỬ LÝ NGÔN NGỮ TỰ NHIÊN**

**Dependency Parsing for**

**English parsing**

*Giảng viên hưỡng dẫn* :

**PGS. TS Lê Anh Cường**

*Người thực hiện - MSSV*:

**Phạm Dương Thành Long – 51603190**

**Ngô Minh Quang – 51603253**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 201**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO CUỐI KÌ MÔN NHẬP MÔN**

**XỬ LÝ NGÔN NGỮ TỰ NHIÊN**

**Dependency parsing for**

**English parsing**

*Giảng viên hưỡng dẫn* :

**PGS. TS Lê Anh Cường**

*Người thực hiện - MSSV*:

**Phạm Dương Thành Long – 51603190**

**Ngô Minh Quang – 51603253**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2019**

LỜI CẢM ƠN

Đầu tiên tôi xin gửi lời cảm ơn chân thành nhất đến các bạn bè trong lớp và thầy Lê Anh Cường nói riêng đã tận tình giúp đỡ chúng tôi trong suốt quá trình thực hiện đề tài cuối kì lần này. Chính điều đó đã thực sự góp phần không nhỏ trong quá trình hoàn thành bài báo cáo cho bài cuối kì lần này. Ngoài ra cũng cảm ơn một số bạn bè khác đã quan tâm đọc bài báo cáo này và về phía nhà trường đã tận tình cung cấp tài liệu để mình hiểu hơn về vấn đề. Tuy bài báo cáo còn sơ sài, do chỉ mới là khái quát cơ bản về bài báo này. Mong các đọc giả đánh giá một cách chân thành và minh bạch để tôi có thể rút kinh nghiệm cho tất cả các bài báo cáo sau, một lần nữa xin cảm ơn.

**ĐỒ ÁN ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Tôi xin cam đoan đây là sản phẩm đồ án của riêng chúng tôi và được sự hướng dẫn của PGS. TS Lê Anh Cường Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình.** Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

TP. Hồ Chí Minh, ngày tháng năm

Tác giả

(ký tên và ghi rõ họ tên)

(Đã ký)

Phạm Dương Thành Long

(Đã ký)

Ngô Minh Quang

PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

Phần xác nhận của GV hướng dẫn

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

Phần đánh giá của GV chấm bài

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

TÓM TẮT

Dependency parsing hay còn gọi là phân tích cú pháp phụ thuộc là một trong những chủ đề nghiên cứu đáng quan tâm cho những ai mới bắt đầu tìm hiểu về NLP hoặc nó còn là một bài toán vẫn đang được nghiên cứu để cải tiến từng ngày ở nhiều phòng nghiên cứu về xử lý ngôn ngữ tự nhiên trên toàn thế giới. Trong bài báo cáo lần này nhóm chúng tôi sẽ làm rõ về các đặc trưng nổi bật trong phân tích phụ thuộc mà mọi người cần nắm được để hiểu sâu hơn về nó, các tính chất, các thuật toán từ cổ điển đến hiện đại. Ngoài ra, nhóm sẽ phát triển thêm về cả mảng ứng dụng (nếu đủ thời gian). Bên cạnh đó trong bài báo cáo lần này chúng ta cũng sẽ bàn thêm về cách chúng hoạt động trong các mô hình như dịch máy hay trích xuất thông tin từ dữ liệu. Quay lại, trong bài viết này sẽ có sử dụng một số tổng hợp từ các bài báo của các trang web có uy tín và các trường đại học nổi tiếng như Stanford hay CoNLL (cho dữ liệu). Phạm vi nghiên cứu lần này sẽ tập trung chính và xây dựng mô hình phân tích cú pháp phụ thuộc cho dữ liệu tiếng anh là chính. Do phạm vi bài báo cáo còn nhỏ nên nhóm sẽ chỉ tập trung vào khai thác những phần nào đáng quan tâm và được xem là quan trọng trong mô hình này. Cảm ơn quý thầy cô và các bạn đã đọc bài báo cáo lần này, nếu có bất kỳ thắc mắc hoặc nhân xét nào có thể liên lạc nhóm chúng tôi để làm rõ hơn về vấn đề. Bây giờ chúng ta sẽ bắt đầu phân tích mô hình sau.

MỤC LỤC

[LỜI CẢM ƠN i](#_Toc40222675)

[PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN iii](#_Toc40222676)

[TÓM TẮT iv](#_Toc40222677)

[MỤC LỤC 1](#_Toc40222678)

[DANH MỤC CÁC HÌNH VẼ VÀ BẢNG BIỂU 3](#_Toc40222679)

[Chương I: Tổng quản về Dependency parsing 5](#_Toc40222680)

[1.1 Depedency parsing là gì ? 5](#_Toc40222681)

[1.1.1 Định nghĩa: 5](#_Toc40222682)

[1.1.2 Một số ứng dụng phổ biến hiện nay: 5](#_Toc40222683)

[1.1.3 Danh sách các loại nhãn phụ thuộc: 6](#_Toc40222684)

[1.2 Tính chất và cách hoạt động của từng nhãn phụ thuộc: 7](#_Toc40222685)

[1.2.1 acomp / xcomp ( adjectival complement): 7](#_Toc40222686)

[1.2.2 advcl (adverbial clause modifier): 8](#_Toc40222687)

[1.2.3 Advmod (adverb modifier): 8](#_Toc40222688)

[1.2.4 amod / Compound (adjectival modifier): 8](#_Toc40222689)

[1.2.5 appos (appositional modifier) : 8](#_Toc40222690)

[1.2.6 aux (auxiliary): 9](#_Toc40222691)

[1.2.7 auxpass (passive auxiliary) 9](#_Toc40222692)

[1.2.8 cc (Coordinating conjunction): 9](#_Toc40222693)

[1.2.9 ccomp (clausal complement): 10](#_Toc40222694)

[1.2.10 conj (conjunct): 10](#_Toc40222695)

[1.2.11 cop (copula): 10](#_Toc40222696)

[1.2.12 csubj /nsubj (clausal subject): 11](#_Toc40222697)

[1.2.13 csubjpass /nsubjpass (clausal subject passsive): 11](#_Toc40222698)

[1.2.14 dep (unspecified dependency): 11](#_Toc40222699)

[1.2.15 det (determiner): 12](#_Toc40222700)

[1.2.16 discourse (discourse element) : 12](#_Toc40222701)

[1.2.17 dobj (direct object) : 12](#_Toc40222702)

[1.2.18 expl (expletive): 13](#_Toc40222703)

[1.2.19 goeswith (goes with): 13](#_Toc40222704)

[1.2.20 iobj (indirect object): 13](#_Toc40222705)

[1.2.21 mark (marker): 13](#_Toc40222706)

[1.2.22 mwe(multi-word expression): 14](#_Toc40222707)

[1.2.23 neg (negation modifier): 14](#_Toc40222708)

[1.2.24 nmod (Nominal modifier): 14](#_Toc40222709)

[1.2.25 nn / compound:nn (noun compound modifier): 15](#_Toc40222710)

[1.2.26 nmod:npmod (noun phrase as adverbial modifier): 15](#_Toc40222711)

[1.2.27 nsubj (nominal subject): 15](#_Toc40222712)

[1.2.28 nsubjpass (Nominal subject passive): 16](#_Toc40222713)

[1.2.29 nummod (Numeric modifier): 16](#_Toc40222714)

[1.2.30 number / compound (element of compound number): 16](#_Toc40222715)

[1.2.31 nmod:poss (possession modifier) 17](#_Toc40222716)

[1.2.32 parataxis (parataxis): 17](#_Toc40222717)

[1.2.33 pcomp (prepositional complement): 18](#_Toc40222718)

[1.2.34 Pobj / case (object of a preposition): 18](#_Toc40222719)

[1.2.35 Preconj / cc:preconj (preconjunct) 18](#_Toc40222720)

[1.2.36 predet / det:predet (predeterminer): 19](#_Toc40222721)

[1.2.37 prep (prepositional modifier): 19](#_Toc40222722)

[1.2.38 prepc (prepositional clausal modifier): 19](#_Toc40222723)

[1.2.39 prt / compound:prt (phrasal verb particle): 20](#_Toc40222724)

[1.2.40 rcmod / acl:relcl (relative clause modifier): 20](#_Toc40222725)

[1.2.41 ref (referent): 20](#_Toc40222726)

[1.2.42 root: 20](#_Toc40222727)

[1.2.43 tmod / nmod:tmod (temporal modifier): 21](#_Toc40222728)

[1.2.44 xcomp (open clausal complement): 21](#_Toc40222729)

[1.3 Các tính chất chính của cây phụ thuộc: 21](#_Toc40222730)

[Chương II : Phương pháp Transition-based 23](#_Toc40222731)

[2.1 Tổng quan về Transition-based: 23](#_Toc40222732)

[2.1.1 Giới thiệu 23](#_Toc40222733)

[2.1.2 Ý tưởng về mô hình: 23](#_Toc40222734)

[2.1.3 Ý nghĩa các thành phần: 24](#_Toc40222735)

[2.1.4 Mô tả thuật toán: 26](#_Toc40222736)

[2.1.4.1 Các tham số trong thuật toán: 26](#_Toc40222737)

[2.1.4.2 Mã giả: 27](#_Toc40222738)

[2.1.5 Ví dụ về bài toán Transition- based: 28](#_Toc40222739)

[2.2 Kết quả thực tế từ việc xây dựng mô hình: 31](#_Toc40222740)

[Huấn luyện mô hình: 32](#_Toc40222741)

[Tham khảo: 34](#_Toc40222742)

DANH MỤC CÁC HÌNH VẼ VÀ BẢNG BIỂU

[Hình 1.1: Một ví dụ về cây cú pháp phụ thuộc theo tiếng anh [2] 5](#_Toc40190491)

[Hình 1.2.1 : ví dụ về acomp / advmod 8](#_Toc40190492)

[Hình 1.2.2: ví dụ về advcl 8](#_Toc40190493)

[Hình 1.2.3: ví dụ về advmod 8](#_Toc40190494)

[Hình 1.2.5: ví dụ về amod / compound 8](#_Toc40190495)

[Hình 1.2.5: ví dụ về appos 9](#_Toc40190496)

[Hình 1.2.6: ví dụ về aux 9](#_Toc40190497)

[Hình 1.2.7: ví dụ về auxpass 9](#_Toc40190498)

[Hình 1.2.8: ví dụ về cc 9](#_Toc40190499)

[Hình 1.2.9: ví dụ về ccomp 10](#_Toc40190500)

[Hình 1.2.10: ví dụ về conj 10](#_Toc40190501)

[Hình 1.2.11: ví dụ về cop 10](#_Toc40190502)

[Hình 1.2.12: ví dụ về csubj 11](#_Toc40190503)

[Hình 1.2.13: ví dụ về dep 11](#_Toc40190504)

[Hình 1.2.14: ví dụ về det 12](#_Toc40190505)

[Hình 1.2.15: ví dụ về discourse 12](#_Toc40190506)

[Hình 1.2.15: ví dụ về dobj / iobj 12](#_Toc40190507)

[Hình 1.2.16: ví dụ về expl 13](#_Toc40190508)

[Hình 1.2.17: ví dụ về goeswith 13](#_Toc40190509)

[Hình 1.2.18: ví dụ về mark 14](#_Toc40190510)

[Hình 1.2.19: ví dụ về mwe 14](#_Toc40190511)

[Hình 1.2.20: ví dụ 1 về neg 14](#_Toc40190512)

[Hình 1.2.21: ví dụ 2 về neg 14](#_Toc40190513)

[Hình 1.2.22: ví dụ về nmod/agent 15](#_Toc40190514)

[Hình 1.2.23: ví dụ về nn 15](#_Toc40190515)

[Hình 1.2.24: ví dụ về nmod:npmod 15](#_Toc40190516)

[Hình 1.2.25 : ví dụ về nsubj 16](#_Toc40190517)

[Hình 1.2.26 : ví dụ về nsubjpass 16](#_Toc40190518)

[Hình 1.2.27 : ví dụ 1 về compound 17](#_Toc40190519)

[Hình 1.2.28 : ví dụ 2 về compound 17](#_Toc40190520)

[Hình 1.2.29 : ví dụ 2 về nmod:poss 17](#_Toc40190521)

[Hình 1.2.30: ví dụ 1 về dep 17](#_Toc40190522)

[Hình 1.2.31: ví dụ 2 về parataxis 18](#_Toc40190523)

[Hình 1.2.32: ví dụ 2 về case 18](#_Toc40190524)

[Hình 1.2.33: ví dụ về cc:preconj 19](#_Toc40190525)

[Hình 1.2.34: ví dụ về det:predet 19](#_Toc40190526)

[Hình 1.2.35: ví dụ về compound:prt 20](#_Toc40190527)

[Hình 1.2.36: ví dụ về acl:relcl 20](#_Toc40190528)

[Hình 1.2.37: ví dụ về ref 20](#_Toc40190529)

[Hình 1.2.38: ví dụ về nmod:tmod 21](#_Toc40190530)

[Hình 1.2.39: ví dụ về xcomp 21](#_Toc40190531)

[Hình 1.3.1: ví dụ về Project tree và non-project tree 22](#_Toc40190532)

[Hình 2.1.1.1: Minh họa về mô hình 23](#_Toc40190533)

[Hình 2.1.3.1: Minh họa về một Buffer mới bắt đầu 24](#_Toc40190534)

[Hình 2.1.3.2: Minh họa về một stack đang xét 2 từ. 24](#_Toc40190535)

[Hình 2.1.3.3: Minh họa về Shift Transition 25](#_Toc40190536)

[Hình 2.1.3.4: Minh họa về left arc 25](#_Toc40190537)

[Hình 2.1.3.5: Minh họa về right arc 26](#_Toc40190538)

[Hình 2.1.3.6: Minh họa về quá trình Reduction ( trong trường hợp h(Wk ) != 0 ) 26](#_Toc40190539)

[Hình 2.1.5.1: Kết quả đầu ra mong muốn của ví dụ 2.1.5 28](#_Toc40190540)

Chương I: Tổng quản về Dependency parsing

1.1 Depedency parsing là gì ?

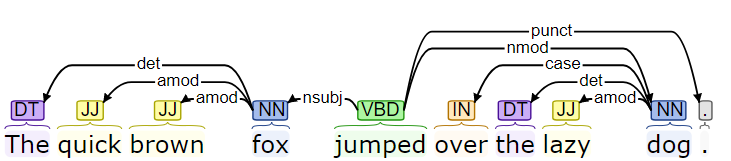
1.1.1 Định nghĩa:

**Depedency parsing** hay còn được gọi là phân tích *cây cú pháp phụ thuộc*, chúng được tạo nên bởi các token trong câu và quan hệ hai ngôi giữa chúng. Một quan hệ phụ thuộc đơn được biểu diễn bởi relation(*governor, depen*dent), trong đó *governer* và *dependent* là *các token*, và *relation* là một kiểu quan hệ phụ thuộc về ngữ pháp. Dạng biểu diễn này về cơ bản là một đồ thị có hướng được gán nhãn gọi là *đồ thị phụ thuộc*, có thể được biến đổi từ cây phân tích cú pháp [[1].](https://ongxuanhong.wordpress.com/2016/02/05/cac-thuat-ngu-trong-xu-ly-ngon-ngu-tu-nhien/)

Các điểm nổi bật:

* Cây phân tích phụ thuộc không phân tích chủ ngữ, vị ngữ, các cụm danh từ, cụm động từ,… thay vì đó nó sẽ phân tích quan hệ phụ thuộc giữa các từ trong câu với nhau.
* Bài toàn trên thường liên quan chặt chẽ đến bài toán Gán nhãn từ loại *( POS Tagging - Part Of Speech Tagging).*
* Chứa đựng nhiều thông tin hơn so với POS phục vụ tốt cho quá trình phân tích ngữ pháp trong câu chặt chẽ hơn.

Ví dụ:



Hình 1.1: Một ví dụ về cây cú pháp phụ thuộc theo tiếng anh [[2]](http://www.linguisticsweb.org/doku.php?id=linguisticsweb:tutorials:linguistics_tutorials:automaticannotation:stanford_core_nlp)

1.1.2 Một số ứng dụng phổ biến hiện nay:

* **Nhận dạng thực thể có tên** *(named-entity recognition - NER*) hay còn gọi là nhận dạng thực thể định danh, xác định thực thể hoặc trích xuất thực thể đều là để chỉ NER.Có nhiệm vụ hỗ trợ việc trích xuất thông tin mà tìm kiếm và phân loại các thành phần nguyên tử trong văn bản vào những loại xác định trước như là tên người, tổ chức, địa điểm, thời gian, số lượng, giá trị tiền tệ, phần trăm v.v[[3]](https://vi.wikipedia.org/wiki/Nh%E1%BA%ADn_d%E1%BA%A1ng_th%E1%BB%B1c_th%E1%BB%83_c%C3%B3_t%C3%AAn)
* **Trích xuất mối quan hệ** *(**relationship extraction):* Dùng để trích xuất các mối quan hệ ngữ nghĩa từ một văn bản. Mối quan hệ được trích xuất thường xảy ra giữa hai hoặc nhiều thực thể của một loại nhất định [[4]](http://nlpprogress.com/english/relationship_extraction.html)
* **Dịch máy** *(machine translation) :* Dùng để thực hiện dịch một ngôn ngữ (gọi là ngôn ngữ nguồn) sang một hoặc nhiều ngôn ngữ khác (gọi là ngôn ngữ đích) một cách tự động, không có sự can thiệp của con người trong quá trình dịch.

1.1.3 Danh sách các loại nhãn phụ thuộc:

Các nhãn loại nhãn phụ thuộc bao gồm:

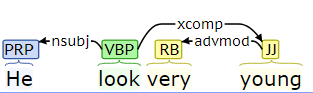
* **acomp / xcomp** *(adjectival complement):* Tính từ bổ ngữ.
* **advcl** *(adverbial clause modifier):* Trạng từ bổ nghĩa mệnh đề.
* **advmod** *(Adverbial modifier):* Trạng từ bổ nghĩa.
* **amod / *Compound*** *(Adjectival modifier):* Tính từ bổ nghĩa.
* **appos** *(appositional modifier)* : Bổ nghĩa định vị / định danh.
* **aux** *(auxiliary):* Trợ động từ.
* **auxpass** *(passive auxiliary):* Trợ động từ thể bị động.
* **cc** *(coordinating conjunction)*: Liên từ.
* **ccomp** *(clausal complement):* Mệnh đề bổ ngữ.
* **conj** *(conjunct)*: Liên ngữ.
* **cop** *(copula):* Hệ từ.
* **csubj** *(clausal subject):* Chủ ngữ mệnh đề. (thay bằng ***nsubj***).
* **csubjpass** *(clausal subject):* Chủ ngữ mệnh đề thể bị động (hiếm dùng).
* **dep** *(unspecified dependency):* Phụ thuộc không xác định.
* **det** *(determiner):* Từ hạn định.
* **discourse** *(discourse element) :* Yếu tố diễn ngôn (hiếm dùng).
* **dobj** *(direct object)* : Tân ngữ trực tiếp.
* **expl** *(expletive):* Cụm từ hỗ trợ.
* **goeswith** *(goes with):* Cụm từ đi liền nhau.
* **iobj** *(direct object)* : Tân ngữ trực tiếp.
* ***mark*** *(marker) :* Đánh dấu.
* **mwe** *(multi-word expression):* Biểu thức đa từ.
* **neg** *(negation modifier)*: Bổ nghĩa phủ định.
* **nmod** *(Nominal modifier):* Danh từ bổ nghĩa.
* **nn / compound:nn** *(noun compound modifier):* Danh từ ghép.
* **nmod:npmod** *(noun phrase as adverbial modifier):* Trạng từ bổ nghĩa.
* **nsubj** *(Nominal subject):* chủ ngữ, chủ thể.
* **nsubjpass***(Nominal subject passive):* chủ ngữ bị động.
* **nummod** *(Numeric modifier):* Số từ bổ nghĩa
* **number /** **compound***(element of compound number):* Quan hệ hỗn hợp.
* **nmod:poss** *(possession modifier):* Cụm từ sỡ hữu.
* **parataxis** *(parataxis):* Đặt cạnh nhau.
* **pcomp** *(prepositional complement):* Giới từ bổ ngữ (hiếm dùng).
* **pobj / case**  *(object of a preposition):* Tân ngữ của một giới từ (pobj thường được thay bằng **case** mối quan hệ đảo lại vị trí 2 từ của pobj).
* **Preconj / cc:preconj** *(preconjunct):* Liên từ trước.
* **Predet / det:predet (**predeterminer): Từ chỉ định.
* **prep***(prepositional modifier):* Giới từ bổ nghĩa *(hiếm dùng).*
* **prepc** *(prepositional clausal modifier):* Cụm giới từ bổ nghĩa *(hiếm dùng hay được thay thế bằng* ***advcl****).*
* **prt***(phrasal verb particle) :* Phân động từ.
* **rcmod / acl:relcl** *(relative clause modifier):* mệnh đề quan hệ bổ nghĩa.
* **ref** *(referent)*: Tham chiếu *(thường dùng trong mệnh đề quan hệ).*
* **root:** nút gốc (vị trí bắt đầu duyệt hay còn được gọi là ***governor***).
* **tmod / nmod:tmod** *(temporal modifier):* Bổ nghĩa cho thời gian
* **iobj** *(indirect object):* Tân ngữ gián tiếp

1.2 Tính chất và cách hoạt động của từng nhãn phụ thuộc:

***1.2.1 acomp / xcomp*** ( adjectival complement):

*Tính từ bổ ngữ:* Dùng để bổ sung của một động từ là một cụm tính từ có chức năng như bổ ngữ (giống như một đối tượng của động từ).

Ex: **acomp**( look, young )

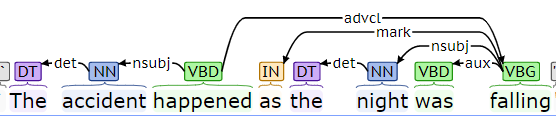


Hình 1.2.1 : ví dụ về acomp / advmod

* + 1. advcl (adverbial clause modifier):

*Trạng từ bổ nghĩa mệnh đề*: Dùng để sửa đổi mệnh đề trạng từ của *VP* hoặc *S* là mệnh đề sửa đổi động từ (mệnh đề tạm thời, hệ quả, mệnh đề câu điều kiện, mệnh đề chỉ mục đích, v.v.).

Ex: **advcl**( happened, falling )

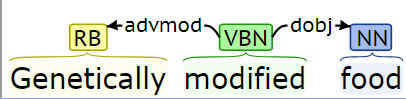


Hình 1.2.2: ví dụ về advcl

1.2.3 Advmod (adverb modifier):

Trạng từ bổ nghĩa: Dùng để sửa đổi trạng từ của một từ là cụm từ chỉ trạng từ *(không phải mệnh đề)* hoặc cụm trạng từ dùng để sửa đổi nghĩa của từ.

Ex: **advmod(** modified, genetically )

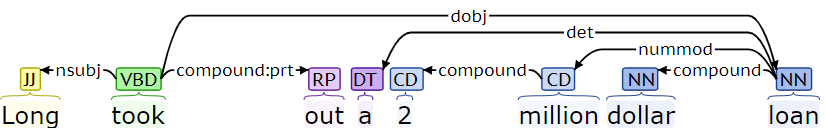


Hình 1.2.3: ví dụ về advmod

1.2.4 amod / Compound (adjectival modifier):

*Tính từ bổ nghĩa:* Dùng để sửa đổi tính từ của một NP*(cụm dành từ)* hay còn được hiểu là bất kỳ cụm tính từ nào dùng để sửa đổi nghĩa của NP.

Ex : ***amod /*Compound**(loan, dollar)

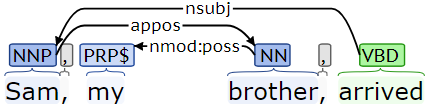
\

Hình 1.2.5: ví dụ về amod / compound

1.2.5 appos (appositional modifier) :

Bổ nghĩa định vị / định danh: Dùng để sửa đổi ứng dụng của NP là NP ngay bên phải NP đầu tiên dùng để xác định hoặc sửa đổi NP đó . Nó bao gồm các ví dụ được ngoặc đơn, cũng như xác định các chữ viết tắt trong một trong các cấu trúc này.

Ex: **appos**( Sam, brother )

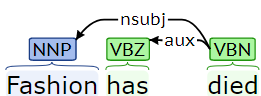


Hình 1.2.5: ví dụ về appos

1.2.6 aux (auxiliary):

*Trợ động từ* được dùng trong mệnh đề là một động từ phụ của mệnh đề (be/ do/ have,..)

Ex: **aux**(died, has)

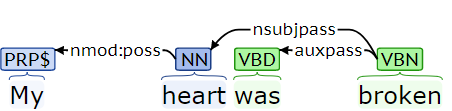


Hình 1.2.6: ví dụ về aux

1.2.7 auxpass (passive auxiliary)

Trợ động từ thể bị động trong một một mệnh đề là một động từ phụ của mệnh đề chứa thông tin bị động.

Ex : **auxpass**( broken, was )

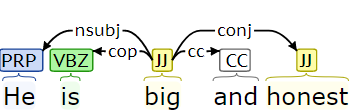


Hình 1.2.7: ví dụ về auxpass

1.2.8 cc (Coordinating conjunction):

*Liên từ:* Là mối quan hệ giữa một liên kết với liên từ đẳng lập trước đó. Một liên từ cũng có thể xuất hiện ở đầu câu. Điều này cũng được gọi là cc, mặc dù không có từ kết hợp trước đó *(ngoại trừ ngầm hoặc trong một câu trước).*

Ex: **cc** (*big*,*and*)

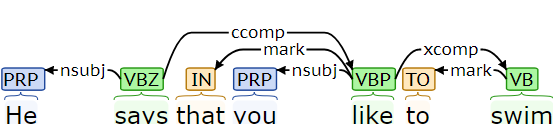


Hình 1.2.8: ví dụ về cc

1.2.9 ccomp (clausal complement):

*Mệnh đề bổ ngữ* của một động từ hoặc tính từ là một mệnh đề phụ thuộc là một đối số cốt lõi. Đó là, nó hoạt động như một đối tượng của động từ, hoặc tính từ.

Ex: **ccomp** *(says,like)*

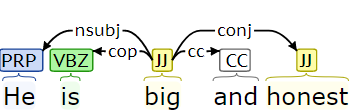


Hình 1.2.9: ví dụ về ccomp

1.2.10 conj (conjunct):

*Liên ngữ*: không phải là một phần của cấu trúc cơ bản của một mệnh đề hoặc câu. Chúng chỉ ra làm thế nào những gì được nói trong câu có liên ngữ nối với những gì được nói trong câu khác hoặc các câu khác. Dùng để thể hiện mối quan hệ giữa 2 từ được kết hợp với nhau ví dụ A and B -> *conj(A,B)*

Ex: **conj** (*big*,*honest*)

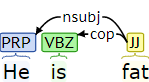


Hình 1.2.10: ví dụ về conj

1.2.11 cop (copula):

*Hệ từ* dùng để thể hiện mối quan hệ của một từ chức năng, được sử dụng để liên kết một chủ đề với một vị từ không lời. Nó thường là một động từ nhưng các từ đồng nghĩa (phi ngôn ngữ) cũng thường xuyên xuất hiện trong các ngôn ngữ thế giới. Các công thức bằng lời nói được gắn thẻ AUX, không phải ĐỘNG TỪ. Các công thức chính được gắn thẻ PRON hoặc DET.

Ex: **Cop** (fat,is)

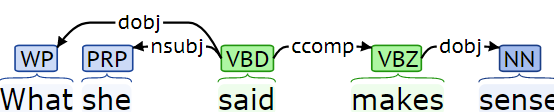


Hình 1.2.11: ví dụ về cop

***1.2.12*** ***csubj /nsubj*** *(clausal subject):*

*Chủ ngữ mệnh đề*: Là chủ ngữ cú pháp mệnh đề của mệnh đề, tức là chủ thể chính nó của nó chính là mệnh đề. Governer của mối quan hệ này không phải luôn là một động từ: khi động từ là một động từ phổ biến, gốc của mệnh đề là bổ ngữ của động từ phổ biến.

Ex: **csubj /nsubj** (said,she)



Hình 1.2.12: ví dụ về csubj

***1.2.13*** ***csubjpass /nsubjpass*** *(clausal subject passsive):*

Chủ ngữ mệnh đề thể bị động là một chủ đề cú pháp mệnh đề của một mệnh đề thụ động. Một chủ ngữ của mệnh đề bị động là một chủ đề cú pháp mệnh đề của một mệnh đề bị động.

Ex: *“That she lied was suspected by everyone”* -> ***csubjpass***(*suspected*, *lied*)

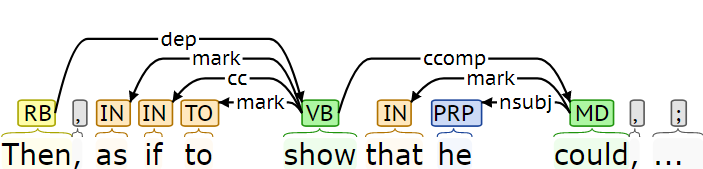
Nhưng hiện nay nhãn này đã bị hạn chế sử dụng hoặc không được phát triển rộng rãi nữa và bị thay thế bằng:

***ccomp***(*lied , suspected*)

1.2.14 dep (unspecified dependency):

Một phụ thuộc không xác định có thể được gắn nhãn là dep khi không thể xác định mối quan hệ chính xác hơn. Điều này có thể là do cấu trúc ngữ pháp kỳ lạ hoặc hạn chế trong phần mềm chuyển đổi hoặc phân tích cú pháp. Việc sử dụng dep nên tránh càng nhiều càng tốt.

Ex: **dep** (*then*,*show)*

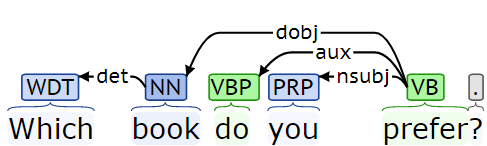


Hình 1.2.13: ví dụ về dep

1.2.15 det (determiner):

Từ hạn định là một từ đứng trước danh từ hoặc một cụm danh từ và trước một tính từ. Do đó mà tự hạn định không thể đứng một mình mà phải có danh từ hoặc một cụm danh từ theo sau.

Ex: **det** (*book*, *which*)

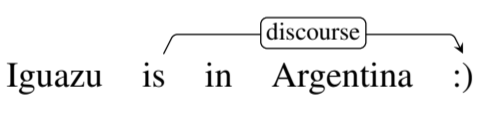


Hình 1.2.14: ví dụ về det

1.2.16 discourse (discourse element) :

*Yếu tố diễn ngôn*,đây là một nhãn phụ thuộc *ít gặp và thường được thay thế bằng* ***advmod****.* Discourseđược sử dụng cho các phần xen kẽ và các phần tử và phần tử diễn ngôn khác *(không liên kết rõ ràng với cấu trúc của câu, ngoại trừ theo cách diễn đạt).* Chúng tôi thường tuân theo các hướng dẫn về những gì mà penTree Treebanks tính là INTJ. Họ định nghĩa điều này bao gồm: sự xen kẽ *( oh , uh-huh , welcome ),* fillers *(Um, ah )* và sự nhấn nhá câu *(well, like, actually, but not you know).*

*Ex : một ví dụ về* discourse(is, :) ) từ stanford denpendency[[6]](https://nlp.stanford.edu/software/dependencies_manual.pdf)



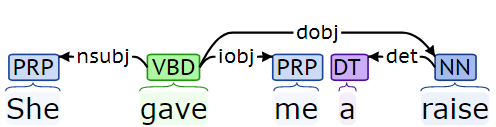
Hình 1.2.15: ví dụ về discourse

1.2.17 dobj (direct object) :

Tân ngữ trực tiếp của VP là một cụm danh từ là đối tượng (đích) của động từ.

Ex : *“She gave me a raise”* -> dobj(gave, raise).

Ex: **dobj**( *gave*,*raise* )

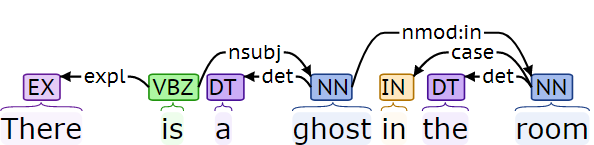


Hình 1.2.15: ví dụ về dobj / iobj

1.2.18 expl (expletive):

Mối quan hệ hỗ trợ này nhầm nắm bắt các đề cử hỗ trợ(expletive) hoặc rờm rà (pleonastic). Đây là những đề cử xuất hiện ở vị trí đối số của một vị ngữ nhưng bản thân chúng không thỏa mãn bất kỳ vai trò ngữ nghĩa nào của vị ngữ. Vị ngữ chính của mệnh đề (động từ hoặc tính từ vị ngữ hoặc danh từ) là governor.

Ex : **expl**( is,there )

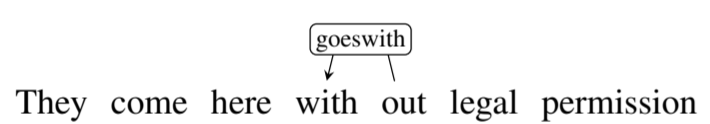


Hình 1.2.16: ví dụ về expl

1.2.19 goeswith (goes with):

*Cụm từ đi liền nhau*: Là mối quan hệ liên kết hai phần của một từ được phân tách trong văn bản không được chỉnh sửa tốt.

*Ex:* ***goeswith*** *(with,out)*[[6]](https://nlp.stanford.edu/software/dependencies_manual.pdf)

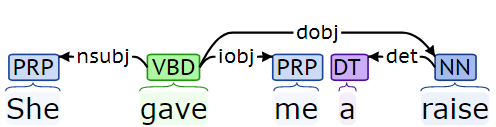


Hình 1.2.17: ví dụ về goeswith

***1.2.20******iobj*** *(indirect object)*:

*Tân ngữ gián tiếp* của VP: Là cụm danh từ là đối tượng (vị cách) của động từ. Trường hợp vị cách là một trường hợp ngữ pháp được sử dụng trong một số ngôn ngữ để chỉ người nhận hoặc người thụ hưởng của một hành động, như trong "Maria Jacobo potum dedit", tiếng Latin có nghĩa là "Maria đã cho Jacob uống".

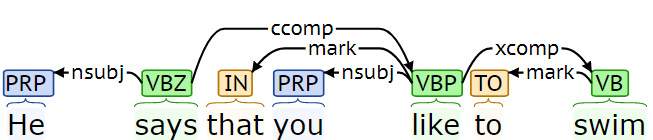
*Ex :* ***iobj*** *(gave, me )*



*Hình 1.2.18: ví dụ về dobj / iobj*

1.2.21 mark (marker):

Một điểm đánh dấu là từ giới thiệu một mệnh đề hữu hạn phụ thuộc vào một mệnh đề khác

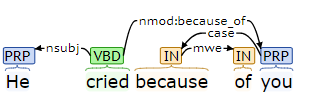


Hình 1.2.18: ví dụ về mark

1.2.22 mwe(multi-word expression):

Mối quan hệ *biểu thức đa từ* (bổ nghĩa) được sử dụng cho một số thành ngữ nhất định hoạt động như một từ đơn chức năng. Ta có một số ví dụ thường gặp*(rather than, as well as, instead of, such as, because of, instead of, in addition to, all but, such as, because of, instead of, due to,v.v..)*

*Ex:* ***mwe****(because,of)*

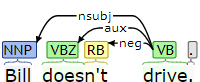


Hình 1.2.19: ví dụ về mwe

1.2.23 ***neg*** *(negation modifier):*

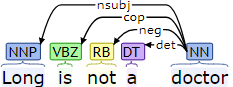
Bổ nghĩa phủ định là mối quan hệ giữa một từ phủ định và từ mà nó sửa đổis

*Ex1:* ***neg*** *(drive, n’t)*



Hình 1.2.20: ví dụ 1 về neg

*Ex2:* ***neg*** (doctor, not)

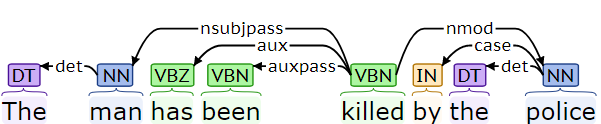


Hình 1.2.21: ví dụ 2 về neg

1.2.24 nmod (Nominal modifier):

*Danh từ bổ nghĩa* được hiểu là sự bổ sung của một động từ bị động, được giới thiệu bởi giới từ thành ngữ bởi phạm lỗi và thực hiện hành động. Mối quan hệ này chỉ xuất hiện trong các phụ thuộc bị sụp đổ, nơi nó có thể thay thế chuẩn bị bằng, khi thích hợp. Nó không xuất hiện trong đầu ra phụ thuộc cơ bản.

Ex: ***nmod*** (killed, police)

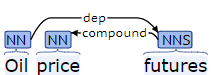


Hình 1.2.22: ví dụ về nmod/agent

1.2.25 nn / compound:nn (noun compound modifier):

*Danh từ ghép (nn )*: Là mội quan hệ kết nối 2 từ trở lên thành một cụm giúp cho một cụm từ đó xác nghĩa hơn, trong quan hệ phụ thuộc nó thường thể hiện dưới dạng ***compound:nn*** .

*Ex :* ***nn*** *(futures, oil)*

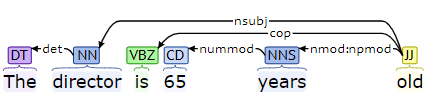


Hình 1.2.23: ví dụ về nn

1.2.26 nmod:npmod (noun phrase as adverbial modifier):

Trạng từ bổ nghĩa là mối quan hệ được xây dựng để biểu thị như một cụm danh từ( NP ), trạng từ bổ nghĩa đứng đầu là các ADJP/ ADVP/ PP và được ước lượng bởi các ADJP/ ADVP. Và thường được đặt trong VP(như là một thành phần bổ nghĩa cho câu).

*Ex:* ***nmod:npmod*** (old, years)

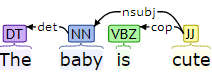


Hình 1.2.24: ví dụ về nmod:npmod

1.2.27 nsubj (nominal subject):

Chủ ngữ, chủ thể là một cụm danh từ là chủ ngữ cú pháp của mệnh đề. Thống đốc của mối quan hệ này có thể không phải luôn luôn là một động từ: khi động từ là một động từ phổ biến, gốc của mệnh đề là bổ ngữ của động từ phổ biến, có thể là một tính từ hoặc danh từ.

Ex :***nsubj***(cute, baby)



Hình 1.2.25 : ví dụ về nsubj

1.2.28 nsubjpass (Nominal subject passive):

chủ ngữ bị động là cụm danh từ hay chủ ngữ cú pháp của mệnh đề dạng bị động.

EX: ***nsubjpass***(defeated, Dole)



Hình 1.2.26 : ví dụ về nsubjpass

1.2.29 nummod (Numeric modifier):

Số từ bổ nghĩa là mối quan hệ dùng để bổ nghĩa cho các danh từ về số lượng hay ước lượng có giá trị cụ thể dạng số hoặc chữ:

Ex:

* *“Sam ate 3 sheep”*

-> **nummod** (sheep, 3)

* *“Sam spent forty dollars”*

-> **nummod** (dollars, 40)

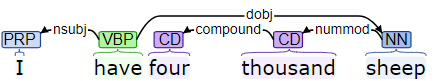
* *“Sam spent $ 40”*

-> **nummod** ($, 40)

1.2.30 number / compound (element of compound number):

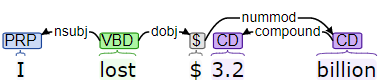
*Mối quan hệ hỗn hợp* là một trong ba mối quan hệ cho các biểu thức đa từ (multiword expressions - MWEs).

*Ex1: “I have four thousand sheep”* -> ***number*** (thousand, four) .



Hình 1.2.27 : ví dụ 1 về compound

*Ex2: “I lost $ 3.2* *billion” ->* ***number*** (billion, 3.2).

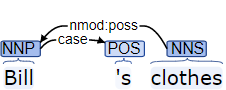


Hình 1.2.28 : ví dụ 2 về compound

1.2.31 nmod:poss (possession modifier)

*Cụm từ sỡ hữu* là mối quan hệ sở hữu giữ giữa người đứng đầu NP và người xác định sở hữu của nó

Ex: “Bill’s clothes” -> ***poss***(clothes, Bill)

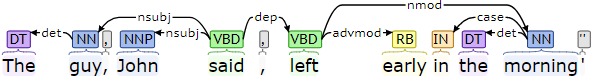


Hình 1.2.29 : ví dụ 2 về nmod:poss

1.2.32 parataxis (parataxis):

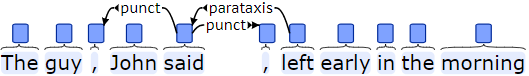
Mối quan hệ *parataxis* (đọc theo tiếng Hy Lạp) có nghĩa là *đặt cạnh bên nhau* là mối quan hệ giữa động từ chính của một mệnh đề và các yếu tố mang tính cảm tính khác, trước và Sau **‘ : ’** hoặc **‘ ; ’.** 2 từ đặt cạnh nhau nhưng không có bất kỳ sự kết hợp rõ ràng, sự phụ thuộc hoặc có quan hệ đối số với từ đầu. tuy nhiên mối quan hệ này thường tuân theo một trật tự nào đó mang tính biểu tượng. Ngoài ra, trong huấn luyện mô hình do giảm số chiều người ta có thể thay thế nó bằng 1 nhãn chung đó là **dep** *( ví dụ 1 hình 1.2.29*

*Ex 1: “The guy , John said , left early in the morning.” ->* ***dep*** *(said,left)*



Hình 1.2.30: ví dụ 1 về dep

*Ex 2: “The guy , John said , left early in the morning.” ->* ***parataxis****: (left,said)*



Hình 1.2.31: ví dụ 2 về parataxis

1.2.33 pcomp (prepositional complement):

*Giới từ bổ ngữ*: Là mối quan hệ thường được sử dụng khi mệnh đề hoặc cụm giới từ (hoặc từ “occasionally”, cụm trạng từ). Giới từ bổ ngữ của giới từ là phần đầu của mệnh đề theo giới từ hoặc phần giới từ của PP (hay thể hiện dưới dạng Mark) :

Ex :***pcomp***(about, missing) ( quan hệ ngược của mark, ko được biểu thị nhiều thường ẩn)

* *“We have no information on whether users are at risk”*

*->* ***pcomp****(on, are).*

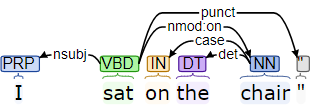
* *“They heard about you missing classes”*

*->* ***pcomp****(about, missing)*

1.2.34 Pobj / case (object of a preposition):

*Tân ngữ của một giới từ*: Là phần đầu của cụm danh từ theo giới từ, hoặc trạng từ “here” và “there”. (Giới từ lần lượt có thể là sửa đổi một danh từ, động từ, v.v.), mặc khác đối tượng có thể được đặt trước giới từ nếu bị kẹt về nghĩa và ngữ pháp.( thông thường đặt sau).

* *Ex1: “I sat on the chair”*
  + ***pobj****(on, chair)*
* *Ex2: “I sat on the chair” ->* ***case****(chair, on)*

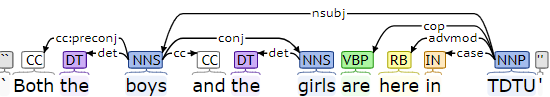


Hình 1.2.32: ví dụ 2 về case

1.2.35 Preconj / cc:preconj (preconjunct)

Một tiền tố là mối quan hệ giữa người đứng đầu NP và một từ xuất hiện ở đầu một từ kết hợp (và nhấn mạnh vào nó), chẳng hạn như: “either”, “both”, “neither”

*Ex:* ***cc: preconj*** *(boys, both)*

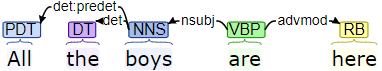


Hình 1.2.33: ví dụ về cc:preconj

1.2.36 predet / det:predet (predeterminer):

Từ chỉ định là mối quan hệ giữa người đứng đầu NP và một từ đi trước và sửa đổi ý nghĩa của bộ xác định NP.

* “All the boys are here” -> ***predet***(boys, all)



Hình 1.2.34: ví dụ về det:predet

1.2.37 prep (prepositional modifier):

Giới từ bổ nghĩa là mối quan hệ của động từ, tính từ hoặc danh từ là bất kỳ cụm từ giới từ nào dùng để sửa đổi nghĩa của động từ, tính từ, danh từ hoặc thậm chí là giới từ khác. Trong biểu diễn bị thu gọn, điều này chỉ được sử dụng cho giới từ với bổ sung là NP.

Ex:

* *“I saw a cat in a hat”*
* ***Prep*** *(cat, in)*
* *“I saw a cat with a telescope”*
* ***Prep*** *(saw, with)*
* *“He is responsible for meals”*
* ***Prep*** *(responsible, for)*

***1.2.38******prepc*** *(prepositional clausal modifier)*:

Giới từ bổ nghĩa mệnh đề là một quan hệ của động từ, tính từ hoặc danh từ là một mệnh đề được giới thiệu bởi một giới từ dùng để sửa đổi nghĩa của động từ, tính từ hoặc danh từ.

Ex:

*“He purchased it without paying a premium”*

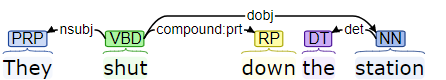
* ***Prepc*** *(purchased, paying)*

1.2.39 prt / compound:prt (phrasal verb particle):

*Phân động từ :* Là mối quan hệ kết hợp giữa một động từ và một phân từ *(particle).* Phân từ này có thể là một trạng từ *(adverb) hoặc* một giới từ *(preposition)* hoặc là cả hai trường hợp trên..

*Ex: “They shut down the station”*

*->* ***compound:prt*** *(shut, down)*

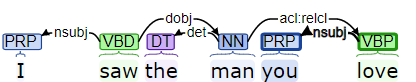


Hình 1.2.35: ví dụ về compound:prt

1.2.40 rcmod / acl:relcl (relative clause modifier):

Quan hệ dùng để liên kết mối quan hệ giữa chủ từ chính và động từ được dùng trong mệnh đề quan hệ( bỏ qua động từ quan hệ mà đi thẳng đến động từ).

*Ex: “I saw the man you love” ->* ***acl:relcl*** *(man, love)*

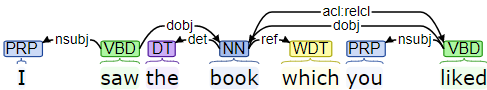


Hình 1.2.36: ví dụ về acl:relcl

1.2.41 ref (referent):

Dùng để biểu hiện quan hệ tham chiếu giữa đại từ quan hệ với chủ từ được sử dụng đứng đằng trước nó

*Ex: “I saw the book which you liked” ->* ***ref*** *(book, which)*



Hình 1.2.37: ví dụ về ref

1.2.42 root:

Được xem là nút gốc của một câu hay vị trí bắt đầu duyệt hay còn được gọi là *governor*. Root có index là 0 và bắt đầu duyệt các từ trong câu bắt đầu tại 1.

*Ex:*

*“I love French fries.”*

* ***root*** *(ROOT, love) .*

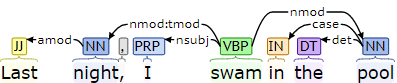
*“Bill is an honest man”*

* ***root*** *(ROOT, man).*

1.2.43 tmod / nmod:tmod (temporal modifier):

*Bổ nghĩa cho thời gian* của VP, NP hoặc ADJP là thành phần cụm danh từ dùng để sửa đổi ý nghĩa của thành phần bằng cách chỉ định thời gian. (Bổ nghĩa tạm thời khác là cụm từ giới từ và được giới thiệu như **prep**.

*Ex:* ***tmod****(swam,night)*

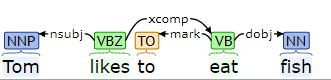


Hình 1.2.38: ví dụ về nmod:tmod

1.2.44 xcomp (open clausal complement):

*Bổ ngữ mệnh đề mở* của một động từ hoặc tính từ là một bổ ngữ dự đoán hoặc mệnh đề không có chủ ngữ riêng.

*Ex: “Tom likes to eat fish” ->* ***xcomp*** *(likes,eat).*



Hình 1.2.39: ví dụ về xcomp

Ngoài ra, Còn một số loại nhãn khác sẽ được cập nhật trong thời gian tới do tính đa dạng về ngữ pháp trong tiếng anh hoặc do phục vụ các một hình khác có độ ổn định cao hơn.

1.3 Các tính chất chính của cây phụ thuộc:

Để làm rõ trong trường hợp này ta sẽ xét một cây cú pháp là một đồ thị với các từ là các ***node*** và các cạnh hay cung *(arc)* dùng để biểu thị quan hệ giữa các node.

***Đồ thị cú pháp phụ thuộc này có 4 tính chất:***

* **Weakly Connected** *(Kết nối yếu):*

***node i***, luôn tồn tại một ***node j*** sao cho có một cạnh nối .

* **Acyclic** *(Không có chu kỳ):*

Nếu tồn tại cạnh , thì ***không thể tồn tại*** một đường đi và ngược lại.

* **Single head** *(1 từ chỉ có duy nhất 1 head):*

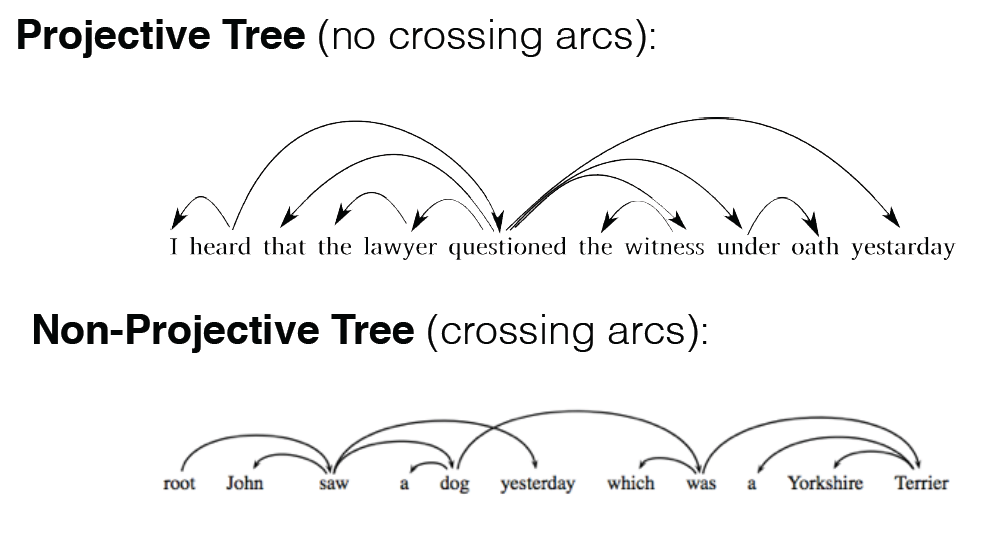
Nếu có cạnh , thì sẽ không có cạnh *(với ).*

* **Projective** *(Tham chiếu):*

Tính tham chiếu áp đặt một ràng buộc bổ sung theo thứ tự từ đầu đến cuối từ các từ ở đầu vào.

○ Nếu tồn tại một cạnh , thì nằm giữa và, luôn có một đường đi

○ ***không có cạnh chéo nhau*** *(cắt nhau)* khi vẽ cây cú pháp tuần tự theo câu



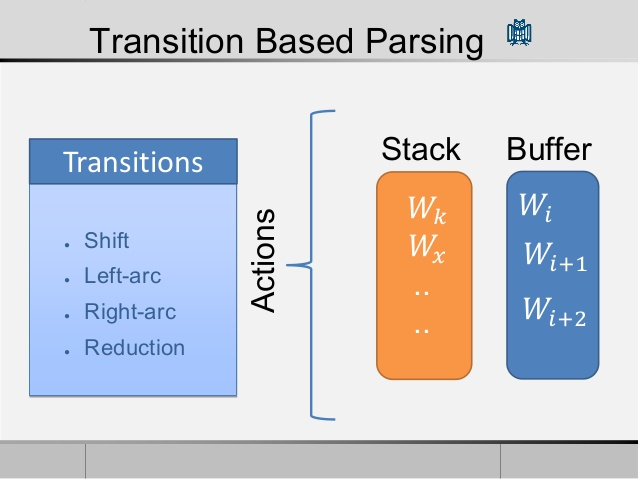
Hình 1.3.1: ví dụ về Project tree và non-project tree

Chương II : Phương pháp Transition-based

2.1 Tổng quan về Transition-based:

2.1.1 Giới thiệu

* ***Transition-based*** là 1 phương pháp phổ biến cho bài toán Dependency Parsing, được nghiên cứu và phát triển bởi các nhà nghiên cứu: *Joakim Nivre, Yoav Goldberg, Yuji Matsumoto,...*



Hình 2.1.1.1: Minh họa về mô hình

* Ý tưởng cơ bản của nó là dựa trên các *Transition bao gồm:*
* Shift
* Reduce
* Left-Arc *(Cung Trái)*
* Right-Arc *(Cung Phải)*
* Khi đọc câu từ trái sang phải, mô hình sẽ quyết định sử dụng các transition nào, dãy các transition này giúp xác định được quan hệ phụ thuộc giữa các từ trong câu.
* Mô hình sử dụng ***thuật toán*** ***Nivre***, là một dạng giải thuật ***greedy*** *(giải thuật tham lam)*

2.1.2 Ý tưởng về mô hình:

* Thực hiện duyệt câu từ trái qua phải, lựa chọn transition có khả năng cao nhất và đi tiếp theo transition đấy.
* Cho phép chỉ tìm ra cây thỏa mãn tính chất *projective*
* Luôn được đánh giá cao về mặt ***tốc độ***, chỉ với ***độ phức tạp O(n)*** *(với n là độ dài của câu)*
* **Có nhiều biến thể**: Arc-Eager, Arc-Standard, Hybrid. Trong báo cáo lần này nhóm chúng tôi sẽ tìm hiều về hệ thống ***Arc-Eager Transition***

2.1.3 Ý nghĩa các thành phần:

Ta có các thành phần chính được dùng trong trong thuật toán như sau:

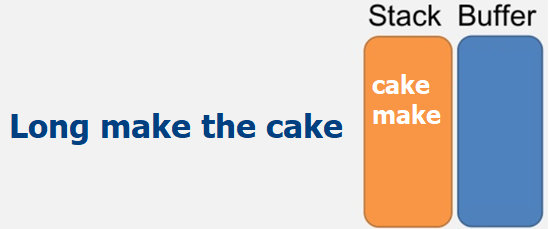
* **Buffer** dùng để chứa các từ chưa được xét, hoặc mới xét đến.

*Ví dụ ta có một câu đầu vào chưa xét:*



Hình 2.1.3.1: Minh họa về một Buffer mới bắt đầu

* **Stack**  dùng để chứa các từ từng được xét và sẽ có thể được xét tiếp.



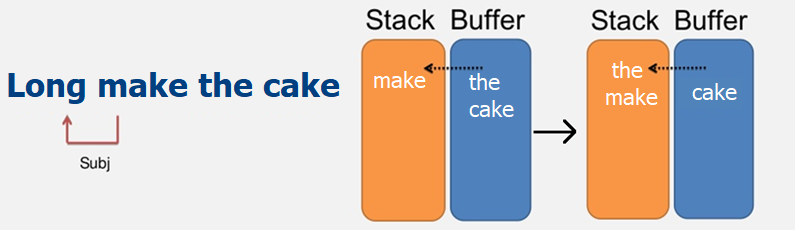
Hình 2.1.3.2: Minh họa về một stack đang xét 2 từ.

* **Tập** dùng để chứa các quan hệ phụ thuộc đã được tìm ra.

Tại mỗi thời điểm, Mô hình sẽ xem xét từ ở đỉnh Stack và từ ở đầu Buffer có quan hệ phụ thuộc gì hay không .

* **Configuration** thường biểu diễn dưới dạng ,*với* **s** *là head và* **b** *dependent*. Dùng để thể hiện 1 trạng thái/cấu hình của hệ thống khi phân tích câu, bao gồm 3 thành phần như trên:
* **Stack**  chứa từ ở đỉnh stack là s.
* **Buffer** uffer B chứa từ b ở đầu tiên (từ tiếp theo xét đến là b).
* **Tập**  các quan hệ phụ thuộc đã tìm ra.
* **Quan hệ phụ thuộc** bao gồm 3 thành phần chính:
* Head là **s**,
* Dependent là **b**,
* Nhãn quan hệ là **lb**.
* **Transition** bao gồm 4 loại chính:
* **Shift** **:**

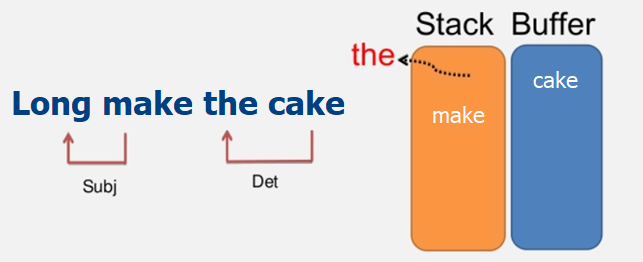
Dùng để chuyển từ ở đầu buffer lên đỉnh của stack và không thêm bất kỳ quan hệ nào.



Hình 2.1.3.3: Minh họa về Shift Transition

* **Left**lb ***(left arc)***

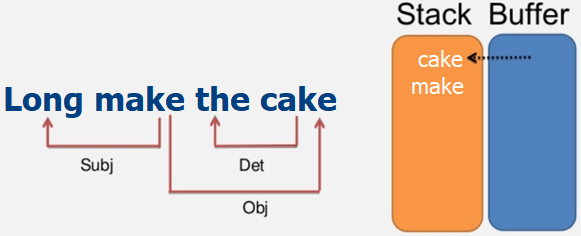
Dùng để bỏ từ ở đỉnh stack ra, giữa nguyên buffer, thêm quan hệ phụ thuộc với với hướng sang trái.



Hình 2.1.3.4: Minh họa về left arc

* **Right**lb ***(right arc)***

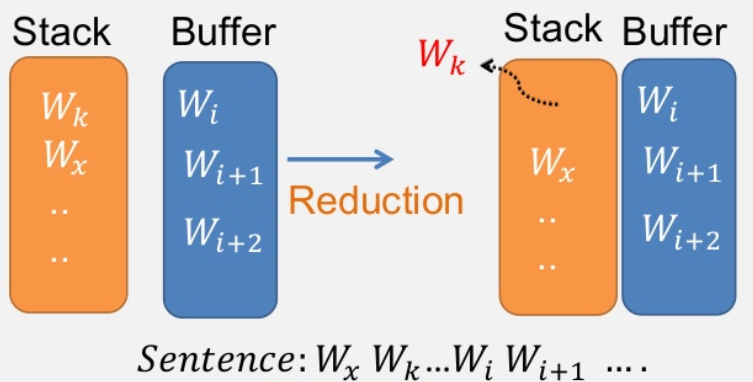
Dùng để thêm từ ở đầu buffer vào đỉnh stack và thêm quan hệ phụ thuộc với với hướng sang phải.



Hình 2.1.3.5: Minh họa về right arc

* **Reduce**

Dùng để bỏ từ ở đỉnh Stack đi và không thêm bất kỳ quan hệ nào.



Hình 2.1.3.6: Minh họa về quá trình Reduction ( trong trường hợp h(Wk ) != 0 )

2.1.4 Mô tả thuật toán:

2.1.4.1 Các tham số trong thuật toán:

* **Input:** W = w1 , w2 , ... , wn , với **w*i*** là từ thứ **i** trong câu
* **Cấu hình khởi tạo** Cinit = trong đó
* : chỉ chứa 1 nút ROOT
* : B = w1 , w2 , ... , wn
* : Tập rỗng
* **Cấu hình kết thúc** Cterminal = trong đó:
* : chỉ chứa 1 nút ROOT
* : Rỗng
* : Tập chứa quan hệ phụ thuộc.
  + - 1. Mã giả:
* Mã giả của thuật toán Transition- based được mô tả như sau:

1. **Input:** *sentence W , parameter-vector w*

4. **return** *Ac*

* Trong đó:
* Hệ thống khởi tạo cấu hình ban đầu theo câu ***W***.
* Kiểm tra xem có phải là cấu hình kết thúc không.
* Tập các transition hợp lệ với cấu hình .
* Transition tiếp theo.
* **w**: Các tham số của mô hình quyết định transition (classifier) và vector này cần được huấn luyện.
* Các đặc trưng được tính tương ứng với cấu hình sau transition
* Giải thích thuật toán:
* Đầu tiên khởi tạo configuration tương ứng với câu đầu vào
* Sau đó, liên tục lựa chọn các transition có khả năng nhất và đi theo transition đấy cho tới khi gặp configuration kết thúc.
* Trả về tập các quan hệ phụ thuộc.

Chú ý: Độ phức tạp là

2.1.5 Ví dụ về bài toán Transition- based:

* Ví dụ ta có câu đầu vào: ***“****We are studying NLP at TDTU****”***
* Bước 1: Xây dựng các biến ban đầu
* Stackvới Root lúc này là ROOT ,
  + Buffer là tất cả các từ còn lại trong câu
  + Tập A *(Tập quan hệ phụ thuộc): rỗng*

|  |  |  |
| --- | --- | --- |
| **TẬP HỢP A** | **STACK** | **BUFFER** |
|  | ***ROOT*** | WE |
|  |  | ARE |
|  |  | STUDYING |
|  |  | NLP |
|  |  | AT |
|  |  | TDTU |

* Bước 2: **Shift** chuyển “**We”** từ *Buffer* sang *Stack:*

|  |  |  |  |
| --- | --- | --- | --- |
| **TẬP HỢP A** | **STACK** | **TRANSITION** | **BUFFER** |
|  | **WE** | (WE) | **~~WE~~** |
|  | ROOT | NLP |
|  |  | AT |
|  |  | TDTU |
|  |  | STUDYING |

* Bước 3:
  + **Left**nsubj xóa “We” ra khỏi Stack
  + ***Thêm*****nsubj**(*studying*, *we*) vào tập hợp A

|  |  |  |  |
| --- | --- | --- | --- |
| **TẬP HỢP A** | **TRANSITION** | **STACK** | **BUFFER** |
| **Nsubj** (*studying*, *we*) |  | **~~WE~~** | ARE |
|  | ROOT | STUDYING |
|  |  | NLP |
|  |  | AT |
|  |  | TDTU |

* Bước 4:
  + **Right**Root thêm “*studying*” Từ *Buffer* vào *Stack*
  + ***Và Thêm*****ROOT**(*root*, *studying*) vào tập hợp A

|  |  |  |  |
| --- | --- | --- | --- |
| **TẬP HỢP A** | **STACK** | **TRANSITION** | **BUFFER** |
| **Nsubj** (*studying*, *we*) | **Studying** |  | **~~STUDYING~~** |
| **ROOT**(*root*, *studying*) | ROOT | NLP |
|  |  | AT |
|  |  | TDTU |
|  |  |  |

* Bước 5: **Shift** chuyển “**are”** từ *Buffer* sang *Stack:*

|  |  |  |  |
| --- | --- | --- | --- |
| **TẬP HỢP A** | **STACK** | **TRANSITION** | **BUFFER** |
| **Nsubj** (*studying*, *we*) | **ARE** | **(are)** | **~~ARE~~** |
| **ROOT**(*root*, *studying*) | Studying | NLP |
|  | ROOT | AT |
|  |  | TDTU |
|  |  | STUDYING |

* Bước 6:
  + **Left**aux xóa “are” ra khỏi Stack
  + ***Thêm***(**aux**, *studying*, *are*) vào tập hợp A

|  |  |  |  |
| --- | --- | --- | --- |
| **TẬP HỢP A** | **TRANSITION** | **STACK** | **BUFFER** |
| **Nsubj** (*studying*, *we*) |  | ~~ARE~~ | STUDYING |
| **ROOT**(*root*, *studying*) | Studying | NLP |
| **aux**(*studying*, *are*) | ROOT | AT |
|  |  | TDTU |
|  |  |  |

* Bước 7:
  + **Right**obj “NLP” từ Buffer vào Stack
  + **Thêm** (**obj**, *studying*, *NLP*) vào *tập hợp A.*

|  |  |  |  |
| --- | --- | --- | --- |
| **TẬP HỢP A** | **STACK** | **TRANSITION** | **BUFFER** |
| **Nsubj** (*studying*, *we*) | **NLP** | (NLP) | **~~NLP~~** |
| **ROOT**(*root*, *studying*) | STUDYING | AT |
| **aux**(*studying*, *are*) | ROOT | TDTU |
| **obj**(*studying*, *NLP*) |  |  |
|  |  |  |

* Bước 8: **Shift** thêm **“at”** từ *Buffer* vào *Stack*

|  |  |  |  |
| --- | --- | --- | --- |
| **TẬP HỢP A** | **STACK** | **TRANSITION** | **BUFFER** |
| **Nsubj** (*studying*, *we*) | **AT** | (AT) | **~~AT~~** |
| **ROOT**(*root*, *studying*) | NLP | TDTU |
| **aux**(*studying*, *are*) | STUDYING |  |
| **obj**(*studying*, *NLP*) | ROOT |  |
|  |  |  |

* Bước 9:
  + xóa “**at**” khỏi Stack
  + **Thêm** (**case**, *TDTU*, *at*) vào *tập hợp A.*

|  |  |  |  |
| --- | --- | --- | --- |
| **TẬP HỢP A** | **TRANSITION** | **STACK** | **BUFFER** |
| **Nsubj** (*studying*, *we*) | (AT) | **~~AT~~** | TDTU |
| **ROOT**(*root*, *studying*) | NLP |  |
| **aux**(*studying*, *are*) | STUDYING |  |
| **obj**(*studying*, *NLP*) | ROOT |  |
| **case**(*TDTU*, *at*) |  |  |

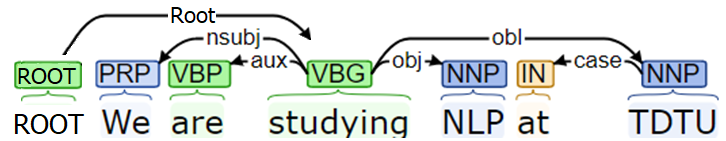
* Bước 10: **Reduce** “NLP” ra khỏi Stack

|  |  |  |  |
| --- | --- | --- | --- |
| **TẬP HỢP A** | **TRANSITION** | **STACK** | **BUFFER** |
| **Nsubj** (*studying*, *we*) | (NLP) | **~~NLP~~** | TDTU |
| **ROOT**(*root*, *studying*) | STUDYING |  |
| **aux**(*studying*, *are*) | ROOT |  |
| **obj**(*studying*, *NLP*) |  |  |
| **case**(*TDTU*, *at*) |  |  |

* Bước 11:
  + **Left**obl “TDTU” từ Buffer vào Stack
  + **Thêm** (**obl**, *studying*, *TDTU*) vào *tập hợp A.*

|  |  |  |  |
| --- | --- | --- | --- |
| * **TẬP HỢP A** | **STACK** | **TRANSITION** | **BUFFER** |
| **Nsubj** (*studying*, *we*) | **TDTU** | (TDTU) |  |
| **Root**(*root*, *studying*) | STUDYING |  |
| **aux**(*studying*, *are*) | ROOT |  |
| **obj**(*studying*, *NLP*) |  |  |
| **case**(*TDTU*, *at*) |  |  |
| **Obl**(*studying*, *TDTU*) |  |  |

* Bước 12:
  + Ta có số quan hệ phụ thuộc bằng **số từ** hay 6 => kết thúc
  + Tổng hợp các quan hệ đã có từ tập hợp A ta được sơ đồ.

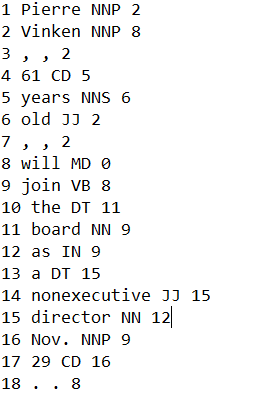


Hình 2.1.5.1: Kết quả đầu ra cuối cùng

* 1. Kết quả thực tế từ việc xây dựng mô hình:

Xử lý tập train: Mỗi câu đầu vào đã được làm sạch và ghi ra một file riêng và mỗi câu sẽ được biểu diễn như sau:

Trong đó ứng với mỗi dòng giá trị thứ nhất là vị trí trong câu, 2-3 là word- postag và 4 là từ mà nó có quan hệ (phụ thuộc)



Hình 2.2.1: Hình ảnh đầu vào đã được làm sạch để training.

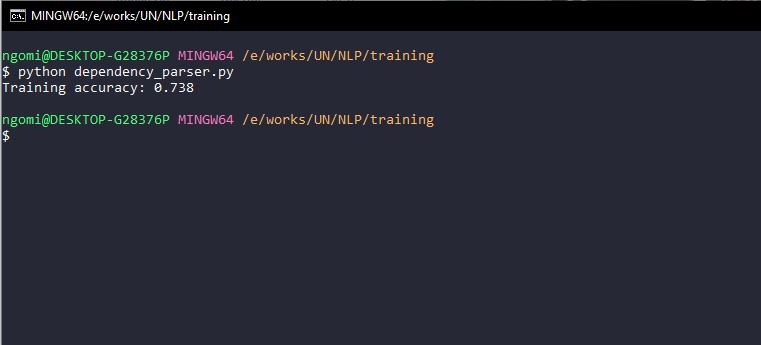
Tiếp theo ta sẽ tổng kết lại số lượng dữ liệu dùng để huấn luyện mô hình:

|  |  |
| --- | --- |
| **Số câu dùng để train** | 3914 |
| **tokens** | 94084 |
| **Số loại POS tags** | 45 |
| **Left-Arcs** | 46061 |
| **Right-Arcs** | 44109 |
| **Root-Arcs** | 3914 |

Hình 2.2.2: Bảng thống kê dữ liệu để training.

**Bên cạnh đó, ta có tổng số Cung của tất cả postag = 349**

Huấn luyện mô hình:



Hình 2.2.3: Độ chính xác khi đánh giá chéo trên mô hình.

Tham khảo:

[1][Các thuật ngữ trong Xử lý ngôn ngữ tự nhiên](https://ongxuanhong.wordpress.com/2016/02/05/cac-thuat-ngu-trong-xu-ly-ngon-ngu-tu-nhien/) – Ông Xuân Hồng.

[2] [Stanford Core NLP](http://www.linguisticsweb.org/doku.php?id=linguisticsweb:tutorials:linguistics_tutorials:automaticannotation:stanford_core_nlp) – by linguistics website

[3] [Nhận dạng thực thể có tên](https://vi.wikipedia.org/wiki/Nh%E1%BA%ADn_d%E1%BA%A1ng_th%E1%BB%B1c_th%E1%BB%83_c%C3%B3_t%C3%AAn) (Theo wikipedia)

[4] [*Relationship extraction*](http://nlpprogress.com/english/relationship_extraction.html) *– by NLP-progress*

[5] [Universal Dependency Relations](https://universaldependencies.org/u/dep/)

[6] [Stanford Denpendency parsing](https://nlp.stanford.edu/software/dependencies_manual.pdf)