HA NOI UNIVERSITY OF SCIENCE AND TECHNOLOGY
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

PHÂN TÍCH VAI NGHĨA

Một số slide được tham khảo từ tutorial của **Scott Wen-tau Yih** & **Kristina Toutanova** (Microsoft Research)

Giới thiệu

Phân tích cú pháp là một bài toán cơ bản trong NLP nhưng kiểu phân tích nào mới là thực sự tốt đối với NLP?

Phân tích cú pháp nhằm thực hiện phân tách câu ra thành các phần có nghĩa hay tìm ra các mối quan hệ có nghĩa mà có thể được sử dụng trong các bài toán tiếp theo về phân tích ngữ nghĩa:

- Gán nhãn vai trò ngữ nghĩa hay còn gọi là phân tích vai nghĩa (chỉ ra ai làm gì cho ai);
- phân tích ngữ nghĩa (chuyển 1 câu thành biểu diễn logic của câu);
- giải quyết nhập nhằng nghĩa từ (chỉ ra các từ trong câu mang ý nghĩa gì);
- xác định ngữ nghĩa hợp thành (tính ý nghĩa của 1 câu dựa trên ý nghĩa các phần của câu).

Trong chương này, chúng ta sẽ tìm hiểu bài toán phân tích vai nghĩa hay gán nhãn vai trò ngữ nghĩa.

Giới thiệu

- Nhiệm vụ chính của *g*án nhãn vai trò ngữ nghĩa (**semantic role labeling - SRL**) là chỉ ra một cách chính xác các quan hệ ngữ nghĩa gì là đúng đắn giữa 1 vị từ và các thành phần kết hợp của nó, trong đó các quan hệ này được lấy ra từ một danh sách đã xác định các vai trò ngữ nghĩa có thể đối với vị từ này.

- Ví dụ:

  [The girl on the swing]*Agent* [whispered]*Pred* to [the boy beside her]*Recipient*

Giới thiệu

- Các vai trò điển hình được sử dụng trong SRL là các nhãn chẳng hạn như Agent, Patient, và Location đối với các thực thể tham gia trong 1 sự kiện, Temporal và Manner để đặc trưng các khía cạnh khác của sự kiện hay các quan hệ tham gia khác.

- Cách tiếp cận theo ngôn ngữ tính toán đối với bài toán SRL đòi hỏi xây dựng một từ điển ngữ nghĩa từ vựng và một bộ sưu tập các câu đã chú thích vai nghĩa.

- 2 kho ngữ liệu được xây dựng dựa trên Ngữ nghĩa khung (frame) có thể sử dụng đ/v bài toán SRL là FrameNet và PropBank.

Các vấn đề đ/v các vai trò ngữ nghĩa

- Rất khó để đưa ra 1 định nghĩa hình thức cho vai trò
- Có các kiểu phân tách vai trò khác nhau tùy ý
- Các giải pháp đ/v vấn đề khó định nghĩa các vai trò ngữ nghĩa:
  - Không chú ý đến nhãn vai trò ngữ nghĩa, chỉ đánh dấu các vai trò/bổ ngữ của các động từ là 0, 1, 2
    - PropBank
  - Xác định các nhãn vai trò ngữ nghĩa đ/v một miền ngữ nghĩa đặc biệt.
    - FrameNet

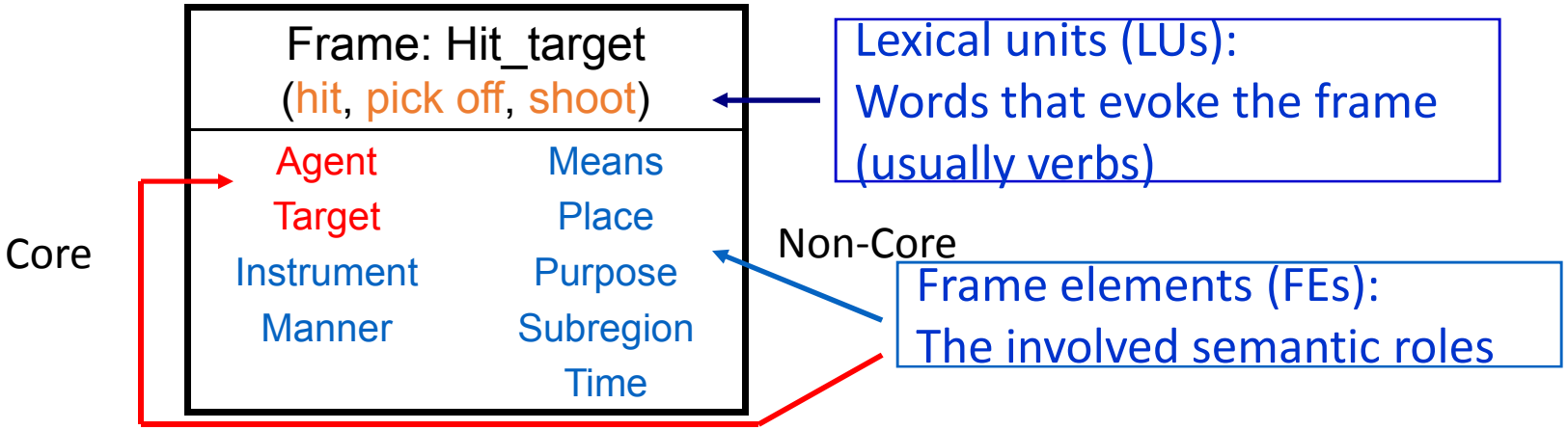
Frame

- **Các ngữ nghĩa khung (frame):**
  - *đề xuất bởi* Fillmore (1976);
  - *khung mô tả 1 trường hợp ở dạng nguyên mẫu*;
  - khung được xác định bởi 1 vị từ (predicate);
  - khung có thể bao gồm một số phần tử khung (arguments; sem. roles).

Frame

- **Các tính chất của ngữ nghĩa khung :**
  - *cung cấp 1 phân tích ngữ nghĩa nông*;
  - *là mức trung gian giữa các vai trò tổng quát và các vai trò đặc biệt theo động từ;*
  - *tổng quát hóa tốt cho các ngôn ngữ khác;*
  - *có thể có lợi cho các ứng dụng NLP khác* (IR, QA).

FrameNet [Fillmore et al. 01]



[$_{\text{Agent}}$ *Kristina*] ***hit*** [$_{\text{Target}}$ *Scott*] [$_{\text{Instrument}}$ *with a baseball*] [$_{\text{Time}}$ *yesterday*].

Frames trong FrameNet

frame(TRANSPORTATION) frame_elements(MOVER(S), MEANS, PATH) scene(MOVER(S) move along PATH by MEANS)
frame(DRIVING) inherit(TRANSPORTATION) frame_elements(DRIVER (=MOVER), VEHICLE (=MEANS), RIDER(S) (=MOVER(S)), CARGO (=MOVER(S))) scenes(DRIVER starts VEHICLE, DRIVER controls VEHICLE, DRIVER stops VEHICLE)
frame(RIDING_1) inherit(TRANSPORTATION) frame_elements(RIDER(S) (=MOVER(S)), VEHICLE (=MEANS)) scenes(RIDER enters VEHICLE, VEHICLE carries RIDER along PATH, RIDER leaves VEHICLE)

Figure 1: A subframe can inherit elements and semantics from its parent

FEG	Annotated Example from BNC
D	[$_D$ Kate] **drove** [$_P$ home] in a stupor.
V, D	A pregnant woman lost her baby after she fainted as she waited for a bus and fell into the path of [$_V$ a lorry] **driven** [$_D$ by her uncle].
D, P	And that was why [$_D$ I] **drove** [$_P$ eastwards along Lake Geneva].
D, R, P	Now [$_D$ Van Cheele] was **driving** [$_R$ his guest] [$_P$ back to the station].
D, V, P	[$_D$ Cumming] had a fascination with most forms of transport, **driving** [$_V$ his Rolls] at high speed [$_P$ around the streets of London].
D+R, P	[$_D$ We] **drive** [$_P$ home along miles of empty freeway].
V, P	Over the next 4 days, [$_V$ the Rolls Royces] will **drive** [$_P$ down to Plymouth], following the route of the railway.

Figure 2: Examples of Frame Element Groups and Annotated Sentences

[Baker, Fillmore, Lowe, 1998]

Các vấn đề đ/v FrameNet

- Các câu mẫu được chọn thủ công
  - Không lựa chọn ngẫu nhiên
  - Không gãn nhãn toàn bộ câu
- Do TreeBank không được sử dụng
  - Không phân tích cú pháp hoàn hảo đ/v câu

Phương pháp luận đối với xây dựng FrameNet

1. Định nghĩa 1 khung (eg DRIVING)
2. Tìm một số câu đối với khung này
3. Chú thích các câu

- Corpora
  - FrameNet I – British National Corpus only
  - FrameNet II – LDC North American Newswire corpora
- Size
  - >8,900 lexical units, >625 frames, >135,000 sentences

http://framenet.icsi.berkeley.edu

Proposition Bank (PropBank) [Palmer et al. 05]

- Dựa trên Penn TreeBank
- Chú thích *mỗi tree* trong Penn TreeBank một cách hệ thống
  - Các thống kê trong corpus này là có ý nghĩa
- Giống FrameNet, dựa trên các lớp động từ của Levin (theo VerbNet)
- Hướng dữ liệu hơn & bottom up
  - Không có mức trừu tượng xa hơn nghĩa động từ
  - Chú thích mỗi động từ xuất hiện trong câu bất kể nó có thuộc khung hay không.

Proposition Bank (PropBank) [Palmer et al. 05]

- Chuyển các câu thành các mệnh đề (propositions)
  - Kristina hit Scott → hit(Kristina,Scott)
- Penn TreeBank → PropBank
  - Thêm 1 tầng ngữ nghĩa trên Penn TreeBank
  - Xác định 1 tập các vai nghĩa đối với mỗi động từ
  - Các vai nghĩa của mỗi động từ được đánh số

…[**A0** the company] to … *offer* [**A1** a 15% to 20% stake] [**A2** to the public]
…[**A0** Sotheby's] … *offered* [**A2** the Dorrance heirs] [**A1** a money-back guarantee]
…[**A1** an amendment] *offered* [**A0** by Rep. Peter DeFazio] …
…[**A2** Subcontractors] will be *offered* [**A1** a settlement] …

Proposition Bank (PropBank)

Xác định tập các vai nghĩa

- Rất khó để xác định được 1 tập các vai nghĩa chung đối với tất cả các kiểu vị từ (verbs).
- PropBank xác định các vai nghĩa và ý nghĩa của chúng đối với mỗi động từ trong frame files.
- Các arguments (core) được đánh nhãn bởi các con số.
  - A0 – Agent; A1 – Patient or Theme
  - Other arguments – no consistent generalizations
- Adjunct-like arguments – ***universal*** đối với tất cả verbs
  - AM-LOC, TMP, EXT, CAU, DIR, PNC, ADV, MNR, NEG, MOD, DIS

Proposition Bank (PropBank) Frame Files

- hit.01 “strike”
  - A0: agent, hitter; A1: thing hit; A2: instrument, thing hit by or with

  [A0 *Kristina*] ***hit*** [A1 *Scott*] [A2 *with a baseball*] *yesterday*.

- look.02 “seeming”
  - A0: seemer; A1: seemed like; A2: seemed to

  [A0 *It*] ***looked*** [A2 *to her] like* [A1 *he deserved this*].

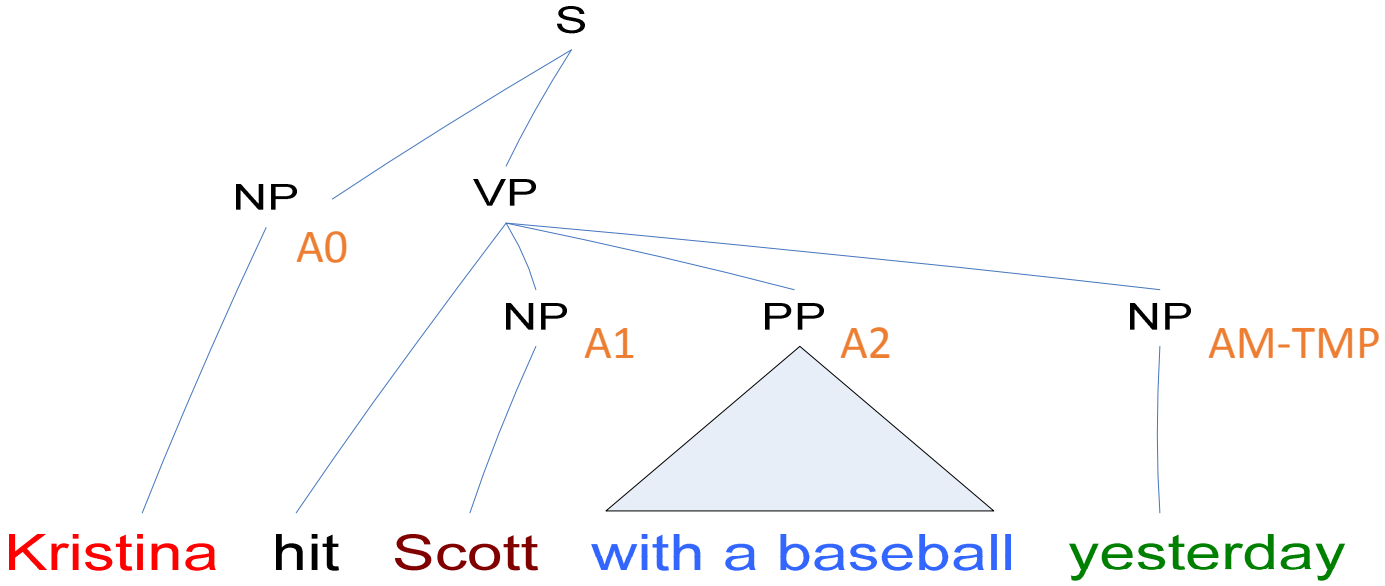
- deserve.01 “deserve”
  - A0: deserving entity; A1: thing deserved; A2: in-exchange-for

  It looked to her like [A0 *he*] ***deserved*** [A1 *this*].

AM-TMP
Time

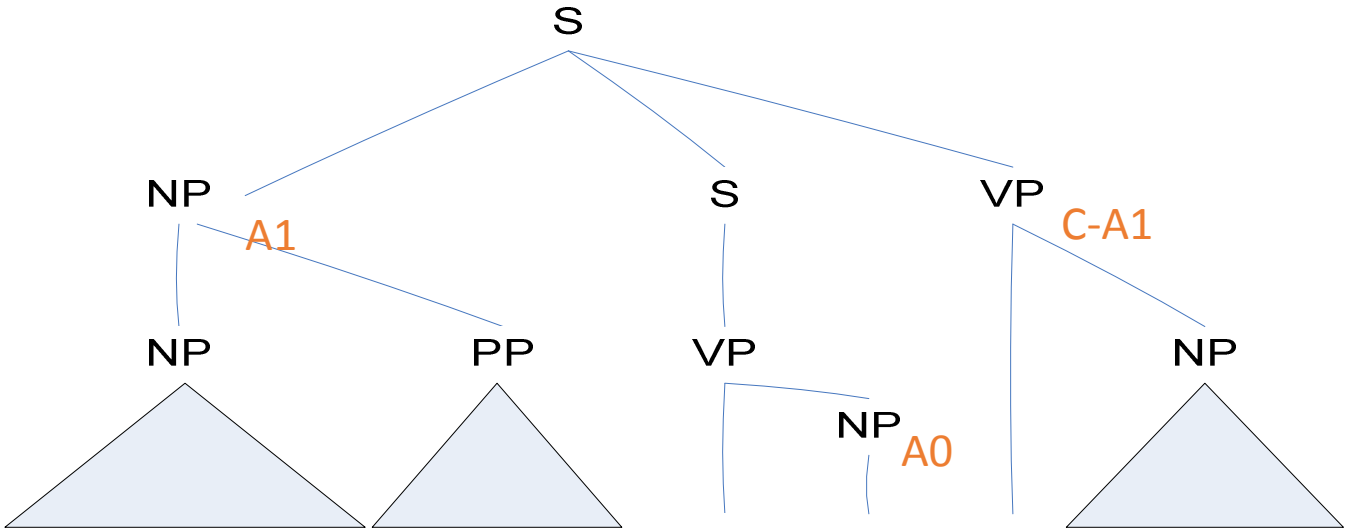
Proposition:
A sentence and a target verb

Proposition Bank (PropBank)
Thêm 1 tầng ngữ nghĩa



[$_{A0}$ *Kristina*] ***hit*** [$_{A1}$ *Scott*] [$_{A2}$ *with a baseball*] [$_{AM-TMP}$ *yesterday*].

Proposition Bank (PropBank)
Thêm 1 tầng ngữ nghĩa– Continued



[$_{A1}$ *The worst thing about him*] ***said*** [$_{A0}$ *Kristina*] [$_{C-A1}$ *is his laziness*].

Một số nghĩa động từ và "framesets" trong propbank

Frameset: **decline.01** "go down incrementally"
Arg1: entity going down
Arg2: amount gone down by, EXT
Arg3: start point
Arg4: end point
Ex: ...[$_{Arg1}$ its net income] *declining* [$_{Arg2\text{-}EXT}$ 42%] [$_{Arg4}$ to $121 million] [$_{ArgM\text{-}TMP}$ in the first 9 months of 1989]. (wsj_0067)

Frameset: **decline.02** "demure, reject"
Arg0: agent
Arg1: rejected thing
Ex: [$_{Arg0}$ A spokesman$_i$] *declined* [$_{Arg1}$ *trace*$_i$ to elaborate] (wsj_0038)

FrameNet vs PropBank -1

FRAMENET ANNOTATION:

[Buyer Chuck] *bought* [Goods a car] [Seller from Jerry] [Payment for $1000].

[Seller Jerry] *sold* [Goods a car] [Buyer to Chuck] [Payment for $1000].

PROPBANK ANNOTATION:

[Arg0 Chuck] *bought* [Arg1 a car] [Arg2 from Jerry] [Arg3 for $1000].

[Arg0 Jerry] *sold* [Arg1 a car] [Arg2 to Chuck] [Arg3 for $1000].

FrameNet vs PropBank -2

FRAMENET ANNOTATION:

[Goods A car] was *bought* [Buyer by Chuck].

[Goods A car] was *sold* [Buyer to Chuck] [Seller by Jerry].

[Buyer Chuck] was *sold* [Goods a car] [Seller by Jerry].

PROPBANK ANNOTATION:

[Arg1 A car] was *bought* [Arg0 by Chuck].

[Arg1 A car] was *sold* [Arg2 to Chuck] [Arg0 by Jerry].

[Arg2 Chuck] was *sold* [Arg1 a car] [Arg0 by Jerry].

Proposition Bank (PropBank)

- Current release (Mar 4, 2005): Proposition Bank I
  - Verb Lexicon: 3,324 frame files
  - Annotation: ~113,000 propositions
    http://www.cis.upenn.edu/~mpalmer/project_pages/ACE.htm
- Alternative format: CoNLL-04,05 shared task
  - Represented in table format
  - Has been used as standard data set for the shared tasks on semantic role labeling

Các vấn đề đ/v PropBank

- Propbank không có danh từ
- Nombank bổ sung đ/v các danh từ
  - NomBank http://nlp.cs.nyu.edu/meyers/NomBank.html
  - Gán nhãn các bổ ngữ xuất hiện với các danh từ trong PropBank

    [A0 *Her*] [REL *gift*] of [A1 *a book*] [A2 *to John*]`

So sánh trích rút thông tin (IE) vs SRL

Characteristic	**IE**	**SRL**
Coverage	narrow	broad
Depth of semantics	shallow	shallow
Directly connected to application	sometimes	no

Tổng quan chung về các hệ thống SRL

- Định nghĩa bài toán SRL
  - Các độ đo đánh giá
- Kiến trúc chung của hệ thống
- Các mô hình học máy
  - Các đặc trưng & các mô hình
  - SRL sử dụng mạng neuron

Các nhiệm vụ con trong SRL

- Nhận diện (Identification): $2^{\{1,2,\ldots,m\}} \mapsto \{NONE, ARG\}$
  - Nhiệm vụ rất khó: tách ra các chuỗi con bổ ngữ từ phần còn lại trong tập có kích thước hàm mũ
  - Thường chỉ có 1 đến 9 (avg. **2.7**) chuỗi con có nhãn ARG còn lại có nhãn NONE đối với 1 vị từ.
- Phân loại (Classification): $2^{\{1,2,\ldots,m\}} \mapsto L \setminus \{NONE\}$
  - Cho 1 tập các chuỗi con có nhãn ARG, quyết định nhãn ngữ nghĩa chính xác
- Gán nhãn vai nghĩa core argument : (dễ hơn)
  - Gán nhãn các cụm với chỉ các nhãn core argument. Các arguments bổ nghĩa (modifier) giả thiết có nhãn NONE.

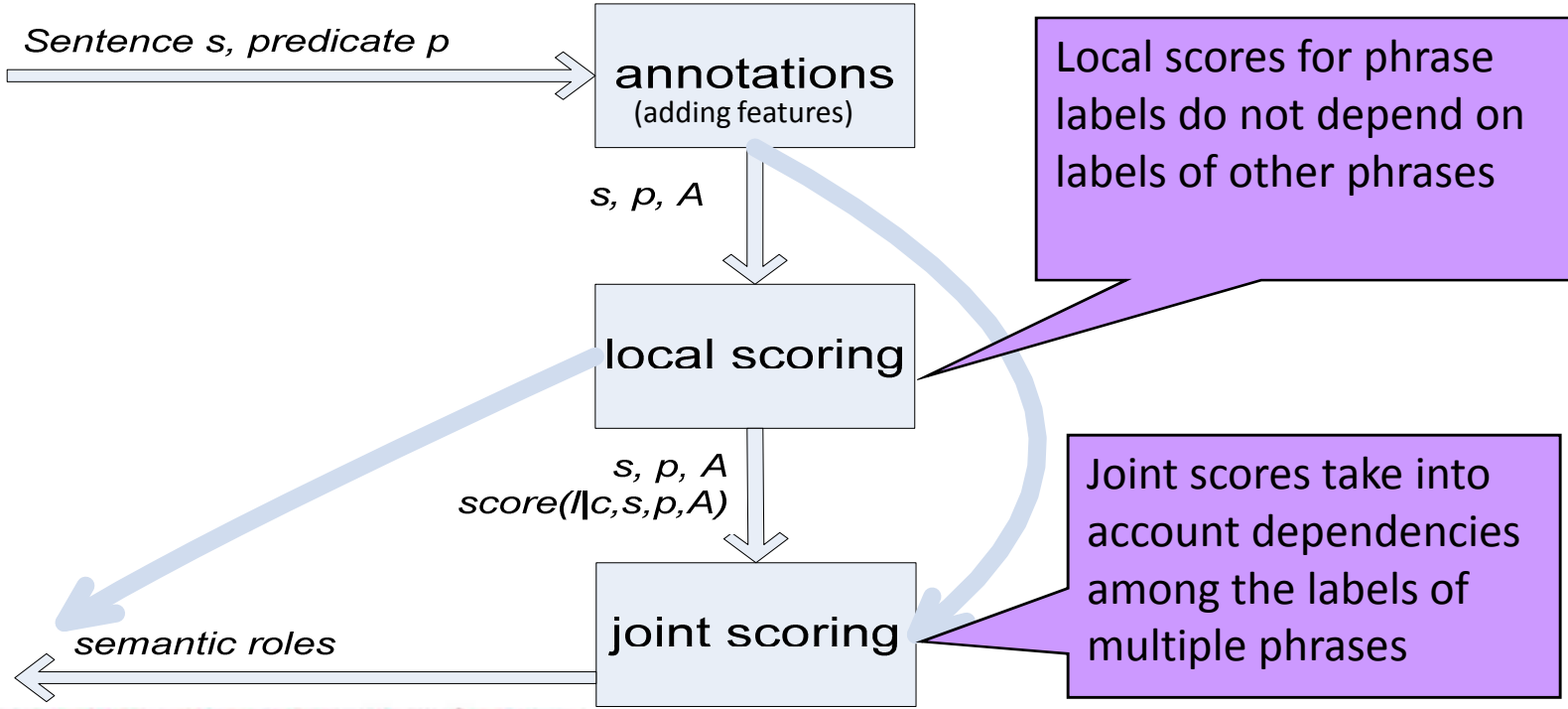
Các độ đo đánh giá

Gán đúng: [$_{A0}$ The queen] **broke** [$_{A1}$ the window] [$_{AM-TMP}$ yesterday]
Dự đoán: [$_{A0}$ The queen] broke the [$_{A1}$ window] [$_{AM-LOC}$ yesterday]

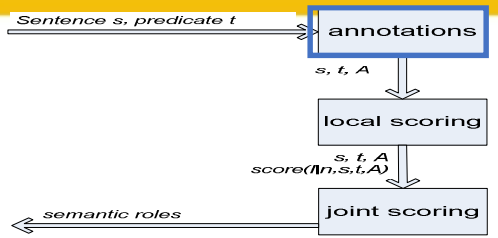
Gán đúng	Dự đoán
{The queen} →A0 {the window} →A1 {yesterday} ->AM-TMP all other → NONE	{The queen} →A0 {window} →A1 {yesterday} ->AM-LOC all other → NONE

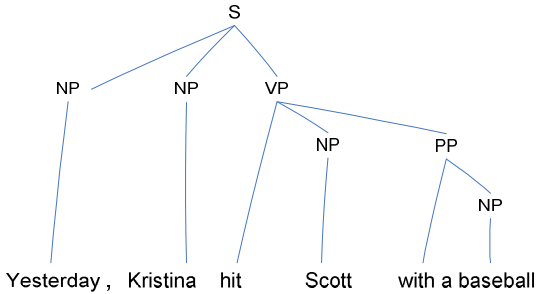
- Precision, Recall, F-Measure {*tp*=1,*fp*=2,*fn*=2} *p*=*r*=*f*=1/3
- Các độ đo đ/v các nhiệm vụ con:
  - Identification (Precision, Recall, F-measure) {tp=2,fp=1,fn=1} *p*=*r*=*f*=2/3
  - Classification (Accuracy) acc = .5 (đánh nhãn các cụm đã nhận diện đúng)
  - Core arguments (Precision, Recall, F-measure) {tp=1,fp=1,fn=1} *p*=*r*=*f*=1/2

Kiến trúc cơ bản chung của 1 hệ thống SRL



Annotations- các chú thích





- Syntactic Parsers
  - Collins', Charniak's (most systems)
    CCG parses ([Gildea & Hockenmaier 03],[Pradhan et al. 05])
    TAG parses ([Chen & Rambow 03])

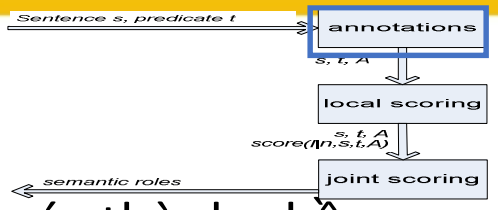
- Shallow parsers

  [$_{NP}$Yesterday] , [$_{NP}$Kristina] [$_{VP}$hit] [$_{NP}$Scott] [$_{PP}$with] [$_{NP}$a baseball].

- Semantic ontologies (WordNet, automatically derived), and named entity classes

  (v) **hit** (cause to move by striking)

WordNet hypernym → **propel, impel** *(*cause to move forward with force*)*



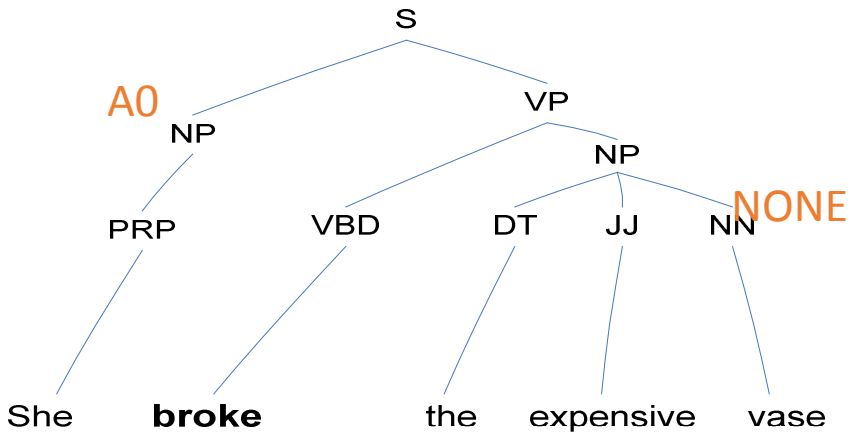
Annotations - Continued

Nói chung, các chuỗi con có nhãn ARG tương ứng với các thành phần cú pháp trong kết quả phân tích cú pháp

- Trong Propbank, 1 cụm ARG tương ứng chính xác với 1 thành phần cú pháp trong cây cú pháp đúng lên đến **95.7**% số các ARG;
  - Khi nhiều hơn 1 thành phần tương ứng với 1 ARG (**4.3**%), các luật đơn gián có thể nối các thành phần con lại với nhau (trong 80% các trường hợp này, [Toutanova 05]);
- Trong Propbank, 1 cụm ARG tương ứng chính xác với 1 thànhphần cú pháp trong cây cú pháp tự động của Charniak với approx **90.0%** số các ARG;
  - Một số (khoảng 30% trường hợp không phù hợp) có thể dễ dàng phục hồi được với các luật đơn giản kết nối các thành phần ([Toutanova 05])
- Trong FrameNet, 1 cụm ARG tương ứng chính xác với 1 thànhphần cú pháp trong cây cú pháp tự động của Collins với **87%** số các ARG.

Đánh nhãn các nút trên cây cú pháp

- Cho 1 cây cú pháp *t*, đánh nhãn các nút (các cụm) trong cây với các nhãn ngữ nghĩa.
- Đối với các ARG không kế tiếp
  - Trong bước hậu xử lý, kết nối 1 số cụm sử dụng các luật đơn giản
  - Sử dụng 1 sơ đồ đánh nhãn mạnh hơn, i.e. C-A0 đ/v sự liên tục của A0



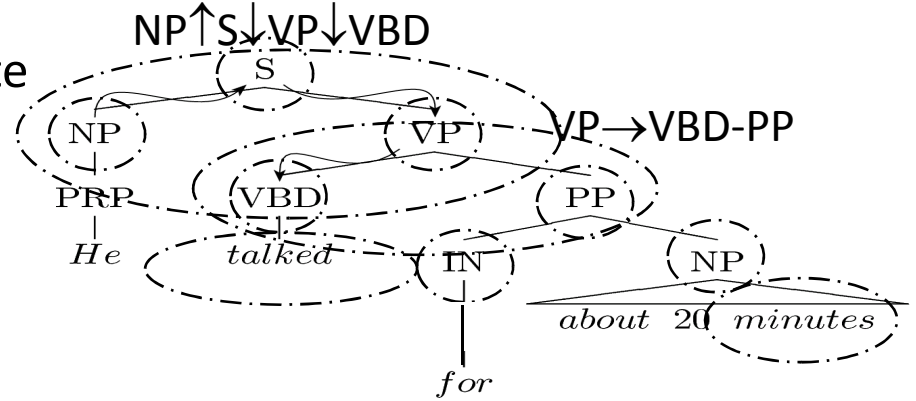
Thuật toán phân tích

- Sử dụng 1 bộ phân tích cú pháp để phân tích cú pháp câu
- Với mỗi vị từ (non-copula verb)
  - Với mỗi nút trong cây cú pháp
    - Trích rút ra 1 vecto đặc trưng ứng với vị từ này
    - Phân loại nút
  - Thực hiện duyệt lần 2 với các thông tin tổng thể

Slide from Sameer Pradhan

Các đặc trưng cơ bản [Gildea & Jurafsky, 2000]

- Predicate (verb)
- Path from constituent to predicate
- Phrase type (syntactic)
- Position (before/after)
- Voice (active/passive)
- Head Word
- Sub-categorization



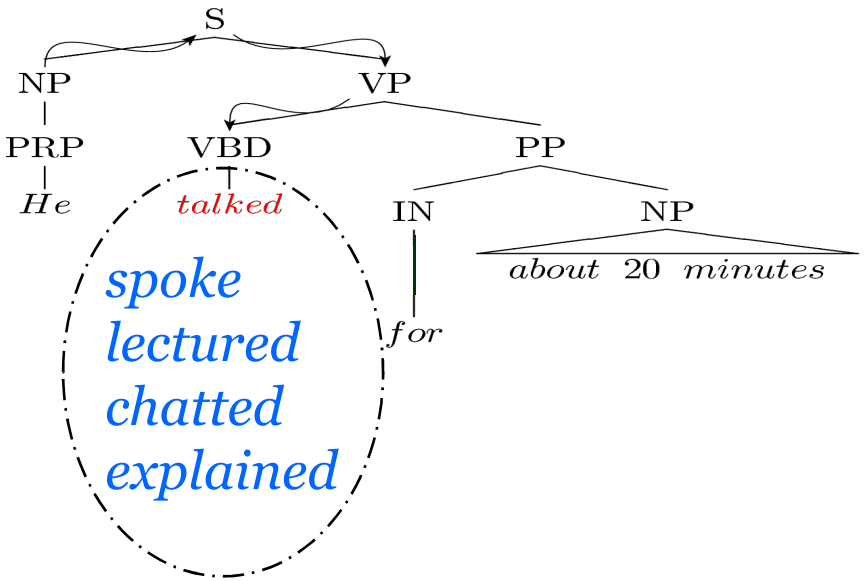
Slide from Sameer Pradhan

Các đặc trưng trong Pradhan et al. (2004)

- Predicate cluster
- Noun head and POS of PP constituent
- Verb sense
- Partial path
- Named entities in constituent (7) [Surdeanu et al., 2003]
- Head word POS [Surdeanu et al., 2003]
- First and last word in constituent and their POS
- Parent and sibling features
- Constituent tree distance
- Ordinal constituent position
- Temporal cue words in constituent
- Previous 2 classifications

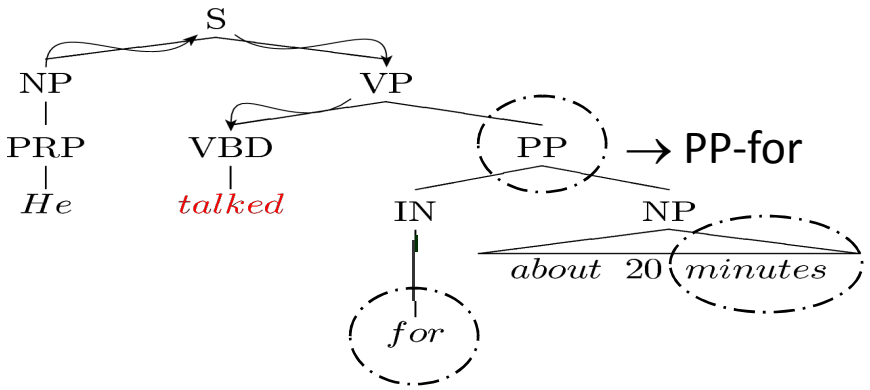
Slide from Sameer Pradhan

Predicate cluster, automatic or WordNet

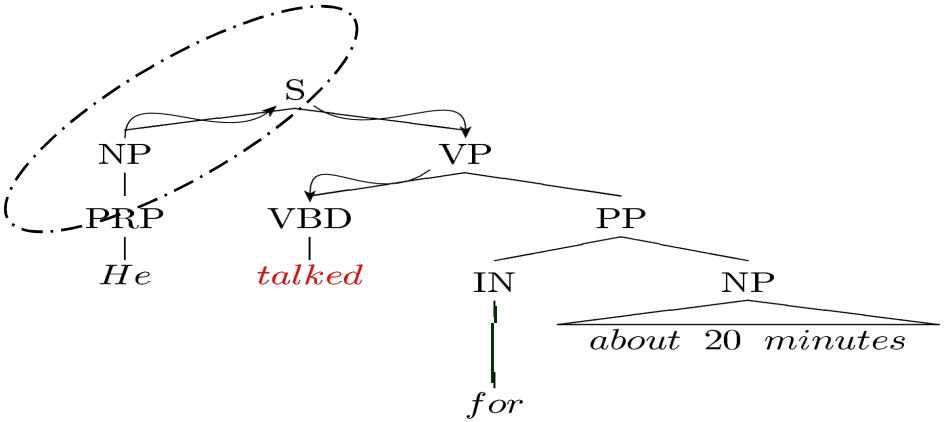


Slide from Sameer Pradhan

Noun Head và POS of PP

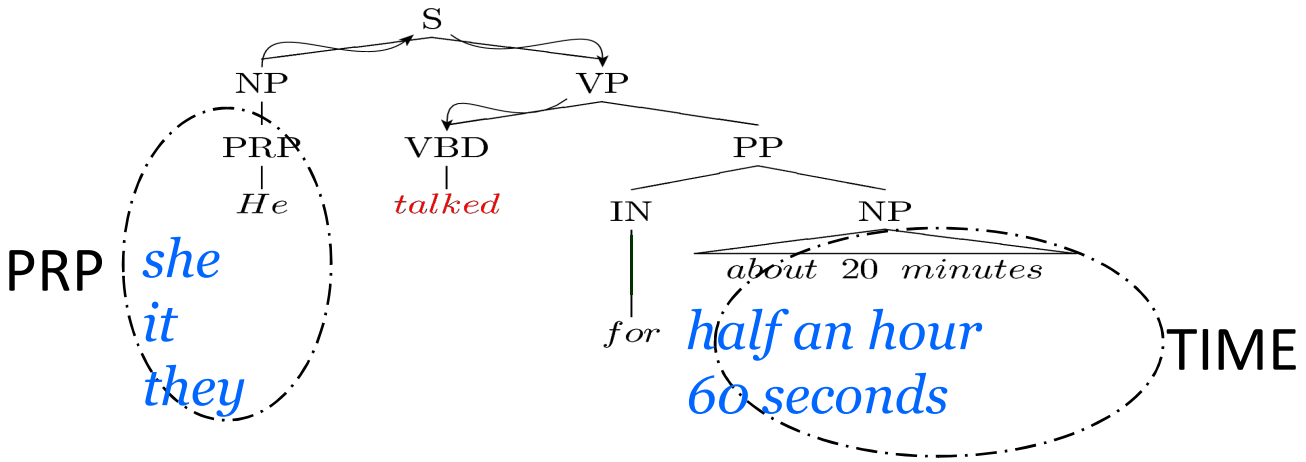


Partial Path

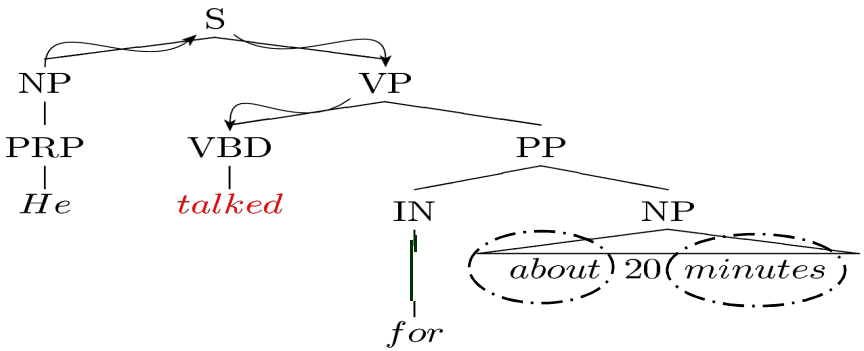


Named Entities and Head Word POS

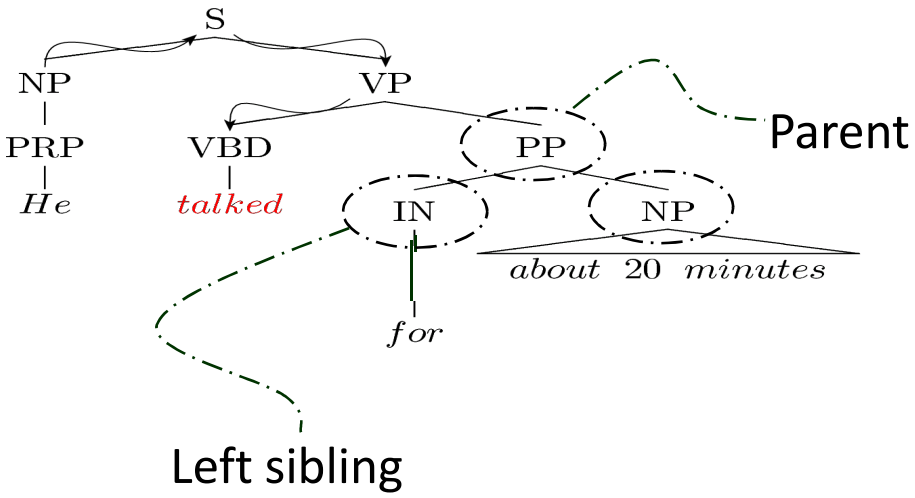
[Surdeanu et al., 2003]



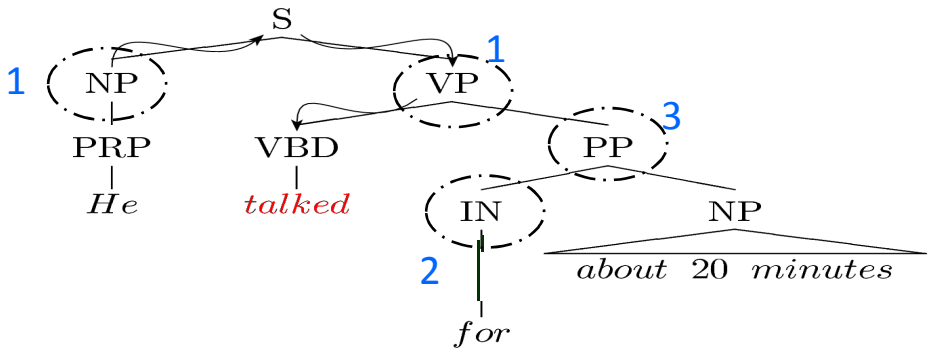
First and Last Word and POS



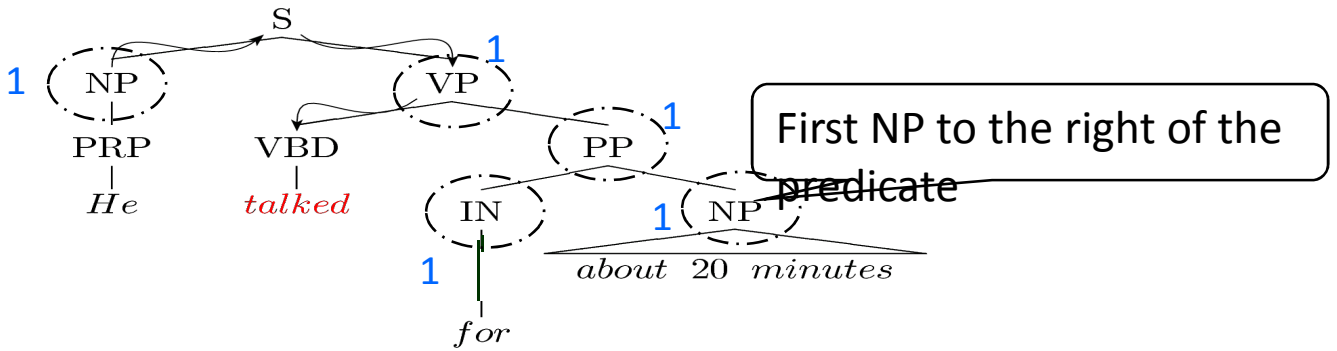
Parent and Sibling features



Constituent tree distance



Ordinal constituent position

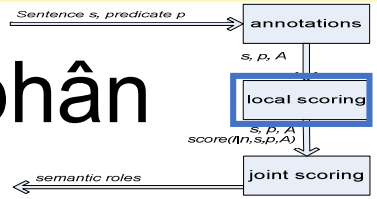


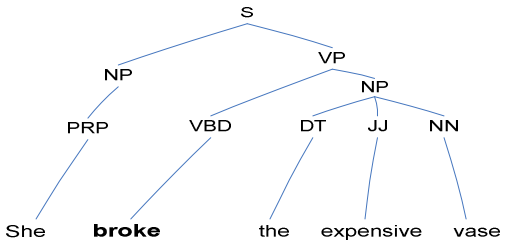
Temporal Cue Words (~50)

time	years;ago
recently	night
days	hour
end	decade
period	late

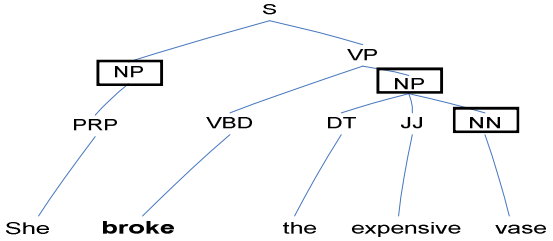
Phân loại nút
(nhận diện nút ARG và phân loại nhãn)

Kết hợp hai mô hình nhận diện và phân loại

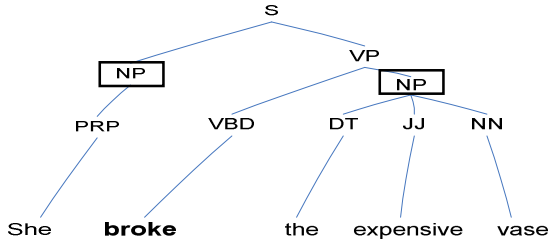




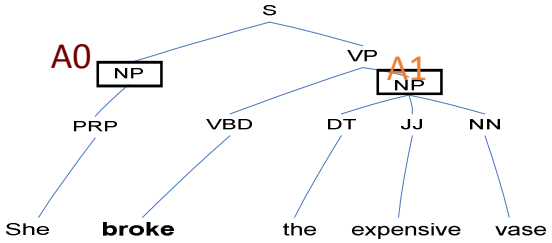
Step 1. *Tỉa cây.*
Dùng 1 bộ lọc thủ công.



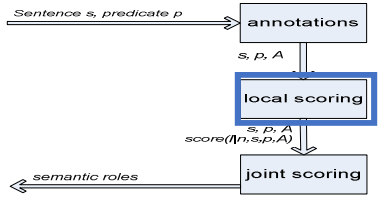
Step 2. *Nhận diện.*
Lọc ra các ứng viên với xác suất cao của NONE



Step 3. *Phân loại.*
Gán 1 trong các nhãn ARG đ/v các nút được chọn (đôi khi có thể là NONE)



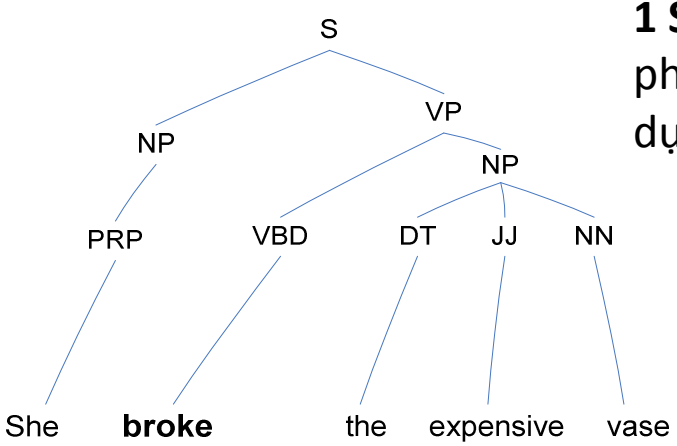
Kết hợp hai mô hình nhận diện và phân loại– Continued



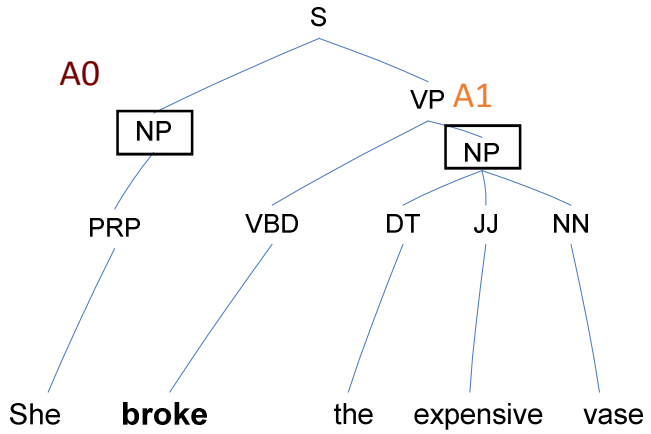
$$P(l|c,t,p) = P_{ID}(Id(l)|\Phi(c,t,p)) * P_{CLS}(l|Id(l),\Phi(c,t,p))$$

or

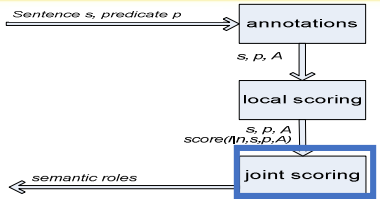
$$P(l|c,t,p) = P(l|\Phi(c,t,p))$$

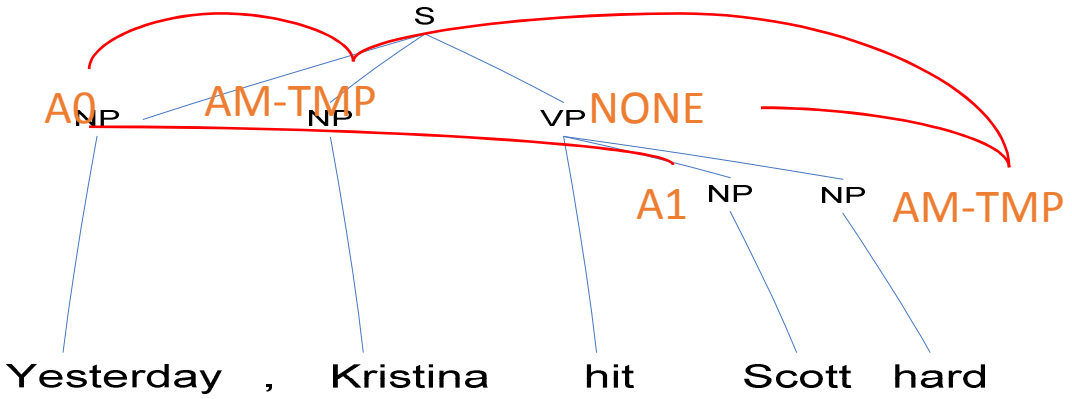


1 Step. Nhận diện và phân loại đồng thời sử dụng $P(l|c,t,p)$



Các mô hình Joint Scoring

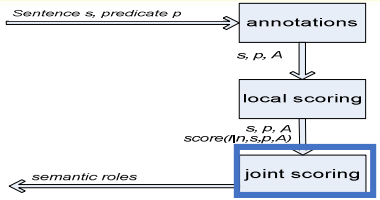




- Các mô hình này tính điểm việc gán nhãn toàn bộ cây (không chỉ các nhãn nút cá thể)
  - Encode some dependencies among the labels of different nodes

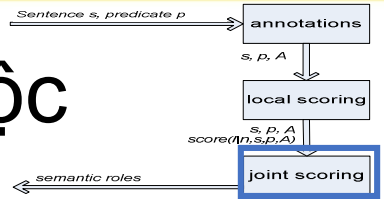
$$P_{JOINT}(l_1,\ldots,l_n|n,t,p) \neq \prod_i P(l_i|n_i,t,p)$$

Kết hợp các mô hình Local và Joint Scoring



- Kết hợp chặt local và joint scoring trong 1 mô hình xác suất đơn và tìm kiếm chính xác [Cohn&Blunsom 05] [Màrquez et al. 05],[Thompson et al. 03]
  - When the joint model makes strong independence assumptions
- **Xếp hạng lại** hay tìm kiếm xấp xỉ để đạt được cách gán nhãn cực đại hóa local và joint score [Gildea&Jurafsky 02] [Pradhan et al. 04] [Toutanova et al. 05]
  - Usually exponential search required to find the exact maximizer
- Tìm kiếm chính xác cách gán tốt nhất mô hình **local thỏa mãn các ràng buộc tổng thể cứng**
  - Using Integer Linear Programming [Punyakanok et al 04,05] (worst case NP-hard)

Joint Scoring: Ép buộc các ràng buộc cứng

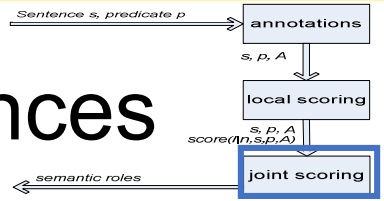


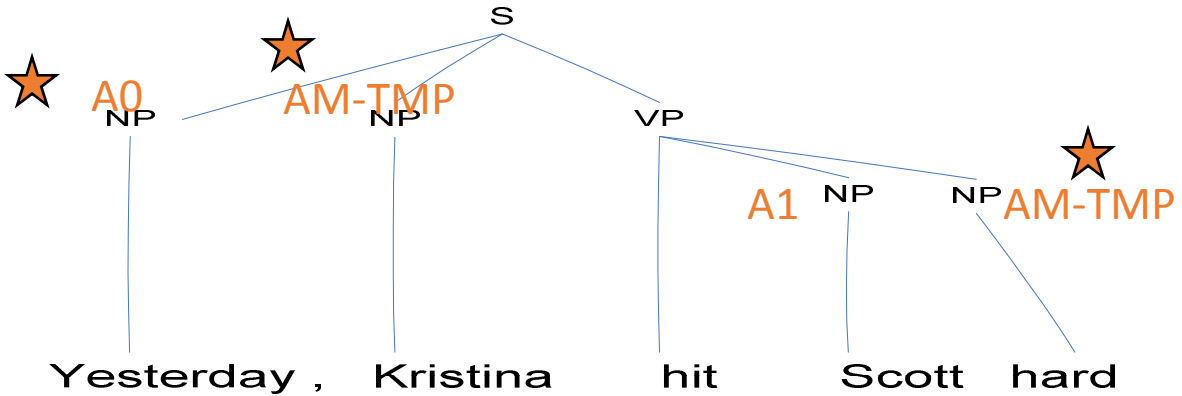
- Ràng buộc 1: Các cụm ARG không bao trùm lên nhau

  *By [$_{A1}$ working [$_{A1}$ hard] , he] **said** , you can achieve a lot.*

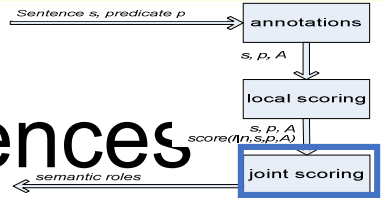
  - Pradhan et al. (04) – tìm kiếm tham lam đ/v 1 tập tốt nhất các ARG không bao trùm lên nhau
  - Toutanova et al. (05) – tìm kiếm chính xác đ/v 1 tập tốt nhất các ARG không bao trùm lên nhau (dynamic programming, linear in the size of the tree)
  - Punyakanok et al. (05) – tìm kiếm chính xác đ/v các ARG không bao trùm lên nhau tốt nhất sử dụng integer linear programming
- Các ràng buộc khác ([Punyakanok et al. 04, 05])
  - core arguments không lặp (good heuristic)
  - Các cụm không bao trùm vị từ
  - (*more later*)

Joint Scoring: Tích hợp Soft Preferences





- Có nhiều xu hướng thống kê đ/v 1 chuỗi các vai trò và các thể hiện cú pháp của chúng
  - Khi cả 2 trước động từ, AM-TMP luôn trước A0
  - Thông thường, không có nhiều temporal modifiers
  - Có thể học tự động nhiều quy tắc khác nữa



Joint Scoring: Tích hợp Soft Preferences

- Gildea and Jurafsky (02) – đánh giá tần suất tương đối trơn của xác suất đa tập phần tử khung.

$$P(\{A0, AM_{TMP}, A1, AM_{TMP}\}|hit)$$

  - Gains relative to local model 59.2 → 62.9 FrameNet automatic parses

- Pradhan et al. (04) – 1 mô hình ngôn ngữ trên các chuỗi nhãn bổ ngữ (with the predicate included)

  - Small gains relative to local model for a baseline system 88.0 → 88.9 on core arguments PropBank correct parses $P(A0, AM_{TMP}, hit, A1, AM_{TMP})$

- Toutanova et al. (05) – 1 mô hình tổng thể dựa trên trên CRFs với 1 tập các đặc trưng chung giàu có của chuỗi các bổ ngữ có nhãn (*more later)*
  - Gains relative to local model on PropBank correct parses 88.4 → 91.2 (24% error reduction); gains on automatic parses 78.2 → 80.0

- Cây CRFs [Cohn & Brunson] đã được sử dụng

Các đặc tính của hệ thống SRL

- Các đặc tính
  - Hầu hết các hệ thống sử dụng tập đặc trưng chuẩn trong Gildea, Pradhan, and Surdeanu
  - *Nhiều đặc trưng là quan trọng đ/v việc xây dựng 1 hệ thống tốt*
- Các phương pháp học
  - SNoW, MaxEnt, AdaBoost, SVM, CRFs, etc.

➢ *Việc lựa chọn các thuật toán học là ít quan trọng*

Các đặc tính của hệ thống SRL– Continued

- Thông tin cú pháp
  - Charniak's parser, Collins' parser, clauser, chunker, etc.
  - Các hệ thống tốt nhất sử dụng Charniak's parser hoặc kết hợp một vài bộ phân tích.

  ➢ *Chất lượng của thông tin cú pháp là quan trọng*

- Kết hợp Hệ thống/Thông tin
  - Greedy, Re-ranking, Stacking, ILP inference

  ➢ *Việc kết hợp các hệ thống hay thông tin cú pháp là chiến lược tốt để giảm ảnh hưởng của thông tin cú pháp không đúng!*

Per Argument Performance

CoNLL-05 Results on WSJ-Test

- Core Arguments (Freq. ~70%)

	Best F_1	Freq.
A0	88.31	25.58%
A1	79.91	35.36%
A2	70.26	8.26%
A3	65.26	1.39%
A4	77.25	1.09%

- Adjuncts (Freq. ~30%)

	Best F_1	Freq.
TMP	78.21	6.86%
ADV	59.73	3.46%
DIS	80.45	2.05%
MNR	59.22	2.67%
LOC	60.99	2.48%
MOD	98.47	3.83%
CAU	64.62	0.50%
NEG	98.91	1.36%

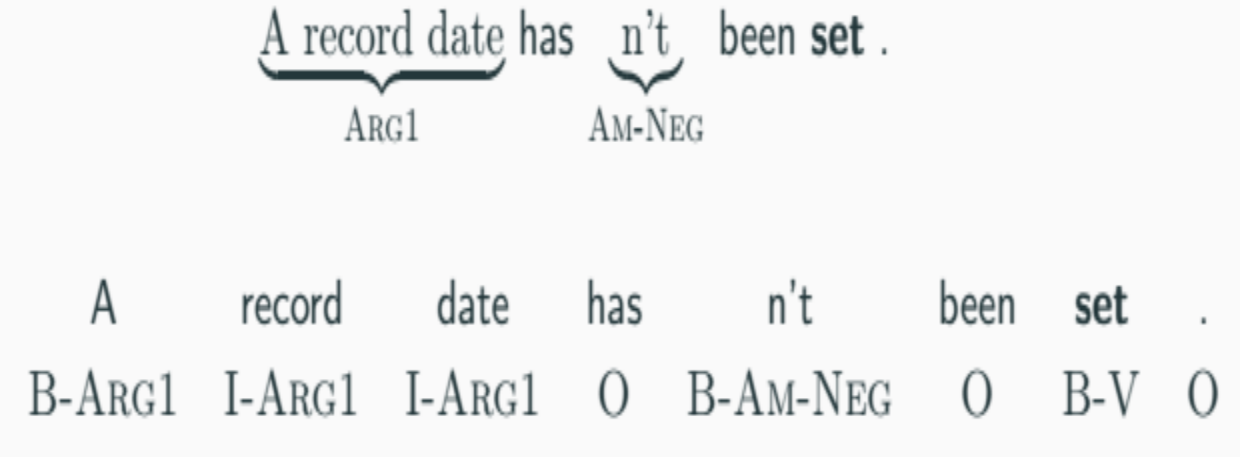
Arguments that need to be improved

Data from Carreras&Màrquez's slides (CoNLL 2005)

SRL sử dụng mạng neuron

Nhận xét: SRL là bài toán gãn nhãn một chuỗi. Do vậy, chúng ta có thể dùng mạng hồi qui (RNNs hoặc LSTMs) đ/v SRL.

SRL sử dụng mạng neuron



SRL sử dụng *deep bi-directional LSTM*

Chúng ta sẽ tìm hiểu 1 hệ thống end-to-end SRL của Zhou & Xu sử dụng *deep bi-directional LSTM (DB-LSTM):*

Các ưu điểm của cách tiếp cận sử dụng deep bi-directional LSTM :

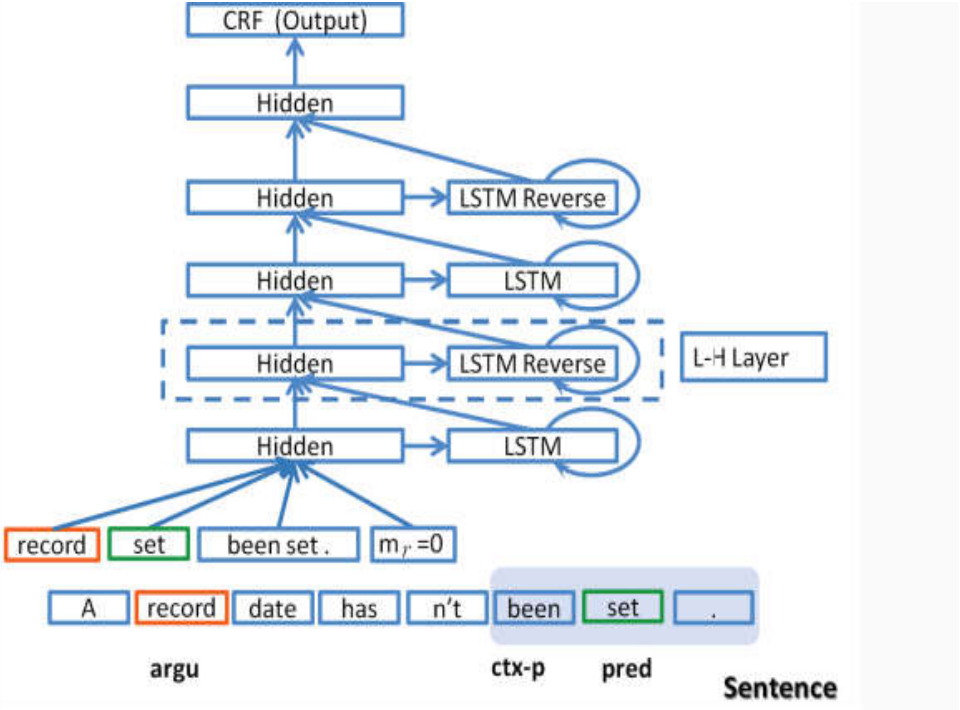
- không sử dụng thông tin cú pháp một cách tường minh;
- *không yêu cầu bước đối sánh phần tử khung riêng rẽ*;
- *không cần các đặc trưng đặc biệt ngôn ngữ thiết kế bởi chuyên gia*;
- *vượt các cách tiếp cận trước đây sử dụng mạng lan truyền tiến*.

Kiến trúc

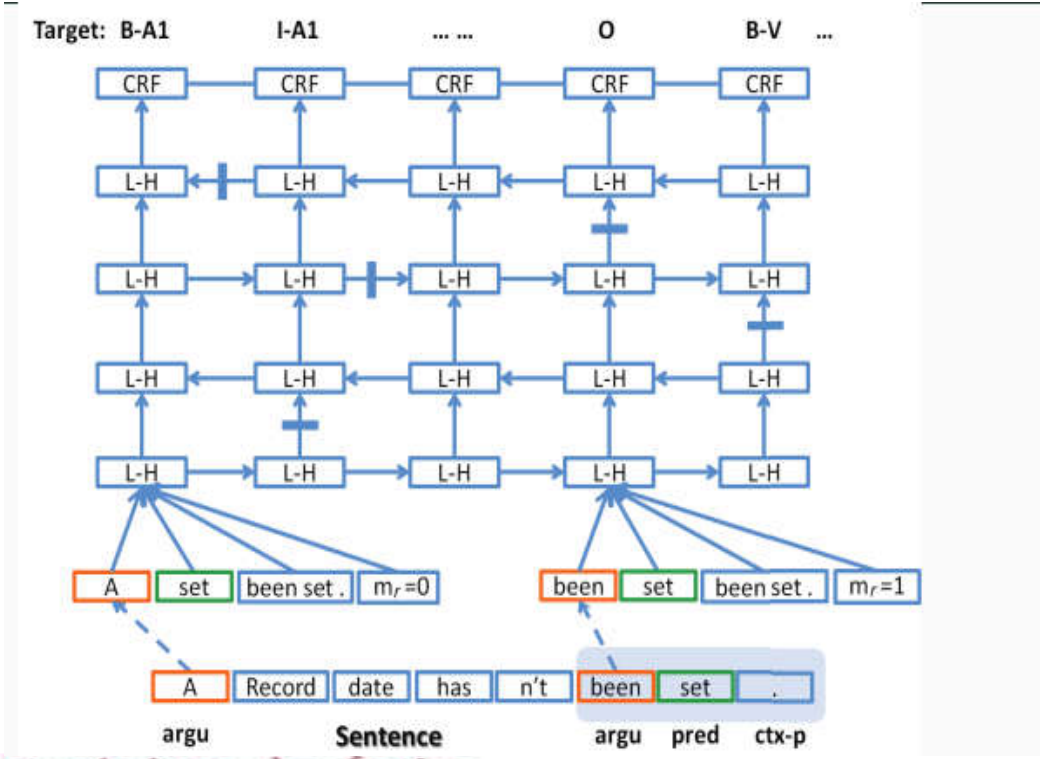
- DB-LSTM được mở rộng từ LSTM chuẩn:

  - LSTM 2 hướng thông thường chứa 2 tầng ẩn, cả hai đều nối đến cùng tầng vào và ra, xử lý cùng chuỗi theo các hướng ngược nhau;
  - *Với* SRL, LSTM 2 hướng được sử dụng một cách khác :
  - 1 tầng LSTM chuẩn xử lý đầu vào theo hướng tiến;
  - *đầu ra của tầng* LSTM này là đầu vào đ/v tầng LSTM khác nhưng theo hướng ngược lại;
  - các cặp tầng LSTM được xếp chồng để đạt được mô hình sâu.

Kiến trúc


CRF (Output)
Hidden
Hidden
LSTM Reverse
Hidden
LSTM
Hidden
LSTM Reverse
L-H Layer
Hidden
LSTM
record
set
been set .
$m_{p'}$=0
A
record
date
has
n't
been
set
.
argu
ctx-p
pred
Sentence

Kiến trúc (unfolded)



Các đặc trưng

- Đầu vào được xử lý từng từ một. Các đặc trưng đầu vào gồm:

  • argument và predicate: argument là từ đang xử lý, predicate là từ nó phụ thuộc vào
  • predicate context (ctx-p): là các từ xung quanh predicate; được sử dụng để phân biệt nhiều thể hiện của cùng predicate;
  • region mark (*mr*): chỉ định liệu argument có ở trong vùng predicate context hay không;
  • nếu chuỗi có *np* predicates thì nó được xử lý *np* lần.

- Đầu ra: nhãn vai trò ngữ nghĩa đ/v cặp predicate/argument sử dụng các thẻ IOB (inside, outside, beginning).

Các đặc trưng
Minh họa với câu ví dụ

Time	Argument	Predicate	ctx-p	m_r	Label
1	A	set	been set .	0	B-A1
2	record	set	been set .	0	I-A1
3	date	set	been set .	0	I-A1
4	has	set	been set .	0	O
5	n't	set	been set .	0	B-AM-NEG
6	been	set	been set .	1	O
7	set	set	been set .	1	B-V
8	.	set	been set .	1	O

Huấn luyện

Các nhúng từ được sử dụng như đầu vào thay cho các từ gốc;

• các nhúng đ/v arguments, predicate, và ctx-p, cũng như *mr* được ghép lại và được sử dụng là đầu vào đ/v DB-LSTM;
• 8 tầng 2 hướng được sử dụng;
• *đầu ra được phân tích qua 1 CRF (*conditional random field); cho phép mô hình hóa các phụ thuộc giữa các nhãn đầu ra;
• mô hình được luyên với standard backprop sử dụng stochastic gradient descent;
.

Các kết quả với CoNLL-2005 Dataset

Embedding	d	ctx-p	m_r	h	F1(dev)	F1
Random	1	1	n	32	47.88	49.44
Random	1	5	n	32	54.63	56.85
Random	1	5	y	32	57.13	58.71
Wikipedia	1	5	y	32	64.48	65.11
Wikipedia	2	5	y	32	72.72	72.56
Wikipedia	4	5	y	32	75.08	75.74
Wikipedia	6	5	y	32	76.94	78.02
Wikipedia	8	5	y	32	77.50	78.28
Wikipedia	8	5	y	64	77.69	79.46
Wikipedia	8	5	y	128	79.10	80.28
Wikipedia	8	5	y	128	79.55	81.07

d: number of LSTM layers; ctx-p: context length; *mr*: region mark used or not; h: hidden layer size. Last row with fine tuning

Các ứng dụng của SRL

- Hỏi đáp tự động
  - Q: When was Napoleon defeated?
  - Look for: [PATIENT Napoleon] [PRED **defeat-synset**] [ARGM-TMP *ANS*]
- Dịch máy

English (SVO)
[AGENT The little boy]
[PRED **kicked**]
[THEME the red ball]
[ARGM-MNR hard]

Farsi (SOV)
[AGENT pesar koocholo] boy-little
[THEME toop germezi] ball-red
[ARGM-MNR moqtam] hard-adverb
[PRED **zaad-e**] hit-past

- Tóm tắt văn bản
  - Predicates và Heads của Roles cho tóm tắt nội dung
- Trích rút thông tin
  - SRL có thể được sử dụng để xây dựng các luật hữu ích đ/v IE

Kết luận

- *Phân tích vai nghĩa (SRL) với mục đích nhận diện ra các thành phần bổ ngữ* (frame elements) xuất hiện trong trường hợp dạng nguyên mẫu (frame) và gán nhãn chúng với các vai nghĩa tương ứng;
- SRL cung cấp cách phân tích ngữ nghĩa nông có lợi cho nhiều ứng dụng NLP khác nhau;
- SRL bao gồm các bước phân tích cú pháp, đối sánh phần tử khung, trích rút đặc trưng, nhận diện/phân loại vai nghĩa;
- SRL cũng có thể coi như bài toán gán nhãn chuỗi và sử dụng bi-directional LSTM luyện trên các nhúng từ để gán nhãn vai nghĩa mà *không cần phân tích cú pháp*, không cần trích rút các đặc trưng thủ công;

Tài liệu tham khảo

1. Carreras, X.and L.Marquez.2005. `Introduction to the CoNLL-2005 Shared Task: Semantic role labeling.In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 152–164, Ann Arbor, MI
2. Cohn, T.and P.Blunsom.2005. Semantic role labelling with tree conditional random fields.In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 169–172, Ann Arbor, MI.
3. Gildea, D.and D.Jurafsky.2002.Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288
4. Marquez, L., P. R. Comas, J. Gim ` enez, and ´N.Catala.2005. Semantic role labeling ` as sequential tagging.In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 193–196, Ann Arbor, MI.
5. Pradhan, S., K.Hacioglu, V.Krugler, W.Ward, J.Martin, and D.Jurafsky.2005a. Support vector learning for semantic argument classification. *Machine Learning*, 60(1):11–39.
6. Toutanova, K., A.Haghighi, and C.Manning. 2005.Joint learning improves semantic role labeling.In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 589–596, Ann Arbor, MI.