BÀI 6: KHAI PHÁ QUAN ĐIỂM (TIẾP)

[4] Phát hiện & gán thực thể

- Trong các đánh giá sản phẩm, thường đã biết được đối tượng được đánh giá
- Tuy nhiên, trên các diễn đàn, cần xác định đối tượng thực thể mà bình luận nhắm tới
  - Tác vụ 1: Phát hiện thực thể trong câu
  - Tác vụ 2: Gán thực thế vào câu (không nêu rõ thực thể được đánh giá)

Giả thiết đồng nhất cảm xúc

VD1:

(1) I bought Camera-A yesterday. (2) I took some pictures in the evening in my living room. (3) The images are very clear. (4) They are definitely better than those from my old Camera-B. (5) The battery is very good too.

VD2: **(4) → Camera-A > Camera-B**

(1) I bought Camera-A yesterday. (2) I took a few pictures in the evening in my living room. (3) The images were very clear. (4) They were definitely better than those from my old Camera-B. (5) The pictures of that camera were blurring for night shots, but for day shots it was ok

Phát biểu bài toán

- Một thread t chứa các post $<p_1, p_2, \ldots, p_n>$
- Một post p chứa các câu $<s_1, s_2, \ldots, s_m>$
- Một câu s chứa các đánh giá về một tập thực thể ε là tập con của tập tất cả các thực thể $E = \{e_1, e2, \ldots\}$
- Một thực thể e có thể xuất hiện tường minh hoặc không tường minh trong một câu s

Phát biểu bài toán (2)

- VD: "*Camera-A looks really pretty. The battery lasts very long*"
- Phần lớn các câu chỉ liên quan đến một thực thể ($|\varepsilon|=1$)
- Câu liên quan đến nhiều thực thể thường là câu so sánh ($|\varepsilon|=2$)
  - "*Camera-A is better than Camera-B*"
- Giả thiết các câu trong một post đều mang ý nghĩa đánh giá đối tượng thực thể (trong thực tế còn có các câu không liên quan, vd chào hỏi)

Phát biểu bài toán (3)

- Cho một tập các threat T trong một lĩnh vực hẹp:
  - Tác vụ 1 - Phát hiện thực thể: Phát hiện tập các thực thể E trong T
  - Tác vụ 2 - Gán thực thể: Gán mỗi câu trong T với một hoặc một vài thực thể trong E

Tác vụ 1 - Phát hiện thực thể

- Phương pháp không giám sát dựa trên khai phá mẫu tuần tự sử dụng một tập thực thể gốc $E^{(0)} = \{e_1, e_2, \ldots, e_n\}$

B1. Chuẩn bị dữ liệu

B2. Khai phá mẫu tuần tự

B3. Trích rút ứng cử viên

B4. Lọc ứng cử viên

B1. Chuẩn bị dữ liệu

- Tìm tất cả các câu chứa các thực thể trong tập gốc; thay thế tên thực thể (chứa một hoặc nhiều từ) bằng tên chung ENTITYXYZ
- Sinh chuỗi bằng cách chọn cửa sổ 5 từ trước và sau thực thể; mỗi phần tử là từ/từ loại

Hiiiiiiiii/NNP SK/NNP -/: ,/, dont/NN be/VB mad/JJ everyone/NN doesnt/NN have/VBP a/DT **n95**/CD phone/NN fetish/NN ducky/JJ

mad/JJ everyone/NN doesnt/NN have/VBP a/DT **ENTITYXYZ** /CD phone/NN fetish/NN ducky/JJ

<{JJ, mad}{NN, everyone}{NN, doesnt}{VBP, have}{DT, a}{CD, ENTITYXYZ}{NN, phone}{NN, fetish} {JJ, ducky}>

B2. Khai phá mẫu tuần tự

- Min support = 0.01
- Các mẫu phải chứa {POS, ENTITYXYZ}
- Mẫu phải có độ dài >= 2
- VD: <{IN}, {DT}, {NNP, ENTITYXYZ }, {is}>

B3. Trích rút ứng cử viên

- Tìm các thực thể khớp với các mẫu sinh ra

The/DT misses/VBZ has/VBZ currently/RB got/VBN **a/DT Nokia/NNP 7390/CD** at/IN the/DT end/NN of/IN the/DT day,/VBG all/DT she/PRP does/VBZ is/VBZ text/NN and/CCmake/VB calls,/NN but/CC the/DT reception/NN is/VBZ terrible,/VBG where/WRB my/PRP$ 6233/CD would/MD get/VB full/JJ bars/NNS hers/PRP would/MD only/RB get/VB 1/CD or/CC 2./CD

<{DT}, {NNP, ENTITYXYZ}, {CD}> ~ a/DT **Nokia**/NNP 7390/CD

<{DT}, {NNP}, {CD, ENTITYXYZ}, {IN}> ~a/DT Nokia/NNP **7390**/CD at/IN

B4. Lọc ứng cử viên

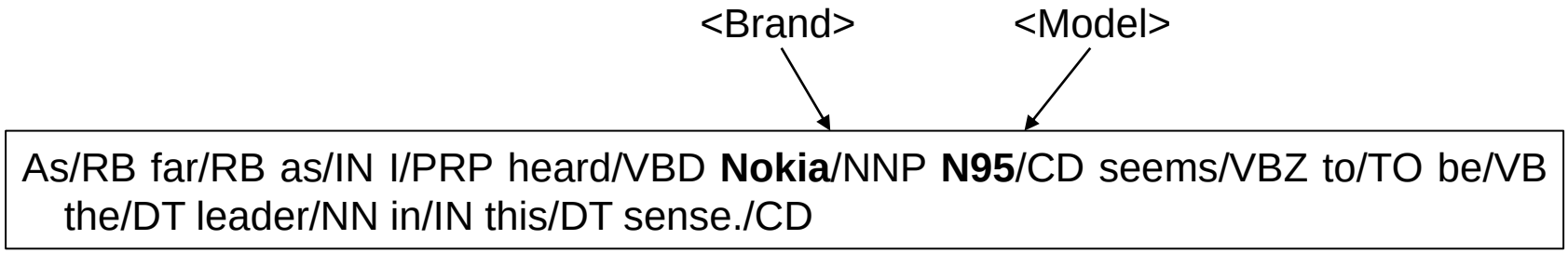
- Loại bỏ các thực thể có POS khác với POS phổ biến nhất với ứng viên này
- VD: ‘accessories’ thường có nhãn NNS nên ‘accessories/CD’ sẽ bị loại

You/PRP can/MD also/RB be/VB sure/JJ it/PRP will/MD work/VB **with/IN all/PDT the/DT Sony/NNP Ericsson/NNP walkman/NN phone/NN accessories/CD**

<{IN}{DT}{CD, ENTITYXYZ}> → accessories **(sai)**

B4. Lọc ứng cử viên (2)

- Sử dụng mẫu <Brand Model> (“Moto Razr V3”) để tìm cặp nhãn hiệu và model
- Sử dụng các mẫu cú pháp để tìm các nhãn hiệu (model) cạnh tranh nhau: A and B; A or B; A vs B; A more than B



Tác vụ 2 - Gán thực thể

- Câu so sánh
  - So sánh hơn: "*Camera-X's battery life is longer than that of Camera-Y*"
  - So sánh bằng: "*Camera-X and Camera-Y are of the same size*"
  - Không so sánh được: "*Camera-X and Camera-Y have different shapes*"
  - So sánh hơn nhất: "*Camera-X's battery life is the longest*"

Đồng nhất cảm xúc

- Giả sử thực thể *e* xuất hiện lần đầu ở câu *s0* và câu tiếp theo là *s1*
- (1) Nếu *s0* là câu bình thường
  - Nếu *s1* là câu bình thường thì nó được gán cho *e*
  - Nếu *s1* là câu so sánh, *e* sẽ được so sánh với một thực thể mới (cần được giới thiệu)
- (2) Nếu *s0* là câu so sánh
  - Nếu *s0* là câu so sánh hơn; *s1* thể hiện cảm xúc tích cực/tiêu cực và không chứa thực thể nào thì nó được gán cho thực thể tốt hơn/kém hơn

Đồng nhất cảm xúc (2)

- Nếu *s*0 là câu so sánh bằng hoặc không so sánh được, do không biết chắc *s*1 đề cập đến thực thể nào, ta gán nó cho thực thể xuất hiện trước *s*0
- Nếu *s*1 là câu so sánh hơn, *s*1 được gán cho thực thể trong *s*1

- (3) Nếu *s*0 là câu so sánh hơn nhất
  - Nếu *s*1 là câu bình thường, ta gán nó cho thực thể tốt nhất được nhắc đến trong *s*0
  - Nếu *s*1 là câu so sánh hơn, *s*1 được gán cho thực thể trong *s*1

Giải thuật

- s_i.entity: Thực thể được nhắc đến trong s_i
- s_i.superiorEntity: thực thể tốt hơn trong câu so sánh hơn
- s_i.inferiorEntity: thực thể kém hơn trong câu so sánh hơn
- ***opinion***(): Hàm xác định cảm xúc trong câu bình thường
- ***compOpinion***(): Hàm xác định cảm xúc trong câu so sánh

```
for each sentence s_i in sequence in a post do
1   If s_i is not a comparative sentence then
2     if s_i contains an explicit entity then
3         s_i.Entity ← the explicit entity of the sentence s_i
4     else    // s_i does not contain an explicit entity
5         if s_{i-1} is not a comparative sentence then
6           s_i.Entity ← s_{i-1}.Entity
7         elseif a superior entity and an inferior entity were
                  discovered in s_{i-1} then
8             opinion(s_i);    // opinion mining
9             if s_i's first clause is a positive clause then
10              s_i.Entity ← s_{i-1}.superiorEntity
11            elseif s_i's first clause is a negative clause then
12              s_i.Entity ← s_{i-1}.inferiorEntity
13            else  s_i.Entity ← s_{i-1}.superiorEntity
14        else s_i.Entity ← s_f.Entity, entities of the last sentence
                       that is not a comparative sentence
15  else      // s_i is a comparative sentence
16    if no entity is mentioned in s_i then
17      s_i.Entity ← s_{i-1}.Entity
18    else s_i.Entity ← {s_{i-1}.Entity} ∪ {entities in s_i};
19      ⟨s_i.inferiorEntity, s_i.superiorEntity⟩ ← compOpinion(s_i)
```

Phân tích cảm xúc

- Phân tích cảm xúc của một câu đối với một thực thể được gán với câu đó dựa trên các bằng chứng:
  - Từ chỉ cảm xúc: great, good, bad, poor; "*the battery of this camera lasts **long***"/ "*This program takes a **long** time to run*"
  - Cụm từ chỉ cảm xúc: "*cost someone an arm and a leg*", "a good deal of"
  - Phủ định: not, "not only ... but also"
  - Mệnh đề 'nhưng': "*The picture quality is great*, ***but** not the battery life*"

Ngôn ngữ đặc tả

```
<rule>      :=  <item> "=>" <action>
<item>      :=  <word> | <word> "[" <type> "]"
<word>      :=  [a-z]+
<type>      :=  JJ | RB | NN | VB | …
<action>    :=  Po | Ne | Neu | Ng | But
```

like[VB] => Po

```
<rule>       :=  <pattern> "=>" <action>
<pattern>    :=  <exp> "+" <target> "+" <exp>
                 | <exp> "+" <target> | <target> "+" <exp>
<exp>        := <element> | <exp> "+" <element>
                 | <exp> "+" <distance> "+" <exp>
                 | <exp> "+" <distance>
                 | <distance> "+" <exp>
                 | !<num> "+" !<item> "+" <exp>
                 | <exp> "+" !<num> "+" !<item>
                 |<exp> "+" !<num> "+" !<item> "+" <exp>
<element>    :=  <item> | item "/" element
<item>       :=  <indicator> | <word>
<indicator>  :=  <indicatorSym>
                 | <indicatorSym> "[" <type> "]"
<target>     :=  <indicator> "[T]" | <word> "[T]"
<indicatorSym> := Po | Ne | Neu | Ng | But
<word>       :=  [a-z]+ | [a-z]+ "[" <type> "]"
<distance>   :=  <num> | <num> - <num>
<num>        :=  0 | [1-9][0-9]*
<action>     :=  <outcome>| !<outcome>
<outcome>    :=  PO | NE | NEU | NG | BUT
<type>       :=  JJ | RB | NN | VB | …
```

VD

The picture quality of this camera is not good, reaction is too slow, but the battery life is long.

*The picture quality is not**[Ng]** good**[Po]**, reaction is too slow**[Neu]**, but**[But]** the battery life is long**[Neu]**.*

too + Neu[JJ][T] => NE

*The picture quality is not[Ng] good[Po], reaction is too slow**[NE]**, but[But] the battery life is long[Neu].*

*The picture quality is not[Ng] good**[Negative]**, reaction is too slow[NE], but[But] the battery life is long[Neu].*

Phân tích câu so sánh

- Câu so sánh khớp một trong các mẫu:
  - a). pronoun + compkey + prodname,
  - b). prodname + compkey + pronoun,
  - c). prodname + compkey + prodname
  - d). pronoun + superkey
  - e). prodname + superkey
  - f). as + JJ + as (ngoại trừ “as long as” và “as far as”)
- Trong đó compkey là từ so sánh, prodname là tên sản phẩm, superkey là từ so sánh hơn

Phân tích câu so sánh (2)

- Các tính từ/trạng ngắn được chuyển sang dạng hơn/hơn nhất bằng cách thêm hậu tố -er/-est (higher/highest)
- Một số trường hợp bất quy tắc: good/better/best
- Các tính từ/trạng từ dài thêm more/most
- Áp dụng quy tắc:
  - more/most + Pos → Positive
  - more/most + Neg → Negative
  - less/least + Pos → Negative
  - less/least + Neg → Positive
- Các từ khác như win, prefer, superior, inferior
  - "*In term of battery life, Camera-X is* ***superior*** *to Camera-Y*"

Đánh giá kết quả

- Tập dữ liệu:
  - HowardForums: Đánh giá phim
  - AVSforums: Plasma/LCD TVs, Projectors and DVD players

Data sets	No. of threads	No. of posts	No. of Product	No. of comparatives	Total no. of sentences
Howard	31	446	171	664	2589
AVS	33	307	180	408	1796
Total	64	753	351	1072	4385

Đánh giá kết quả (2)

- CRF: Phát hiện thực thể dựa trên CRF
- NET: Bộ phát hiện thực thể
- Baseline1: Lấy thực thể cuối của câu trước nếu câu hiện tại không chứa thực thể
- Baseline2: Lấy thực thể đầu tiên của câu trước nếu câu hiện tại không chứa thực thể
- ED (k-com): Các câu so sánh đã biết trước
- ED (unk-com): Cần phát hiện các câu so sánh

Đánh giá kết quả (3)

Table 2: Results of entity discovery

Datasets	CRF		NET		EI (1-3)		EI (1-4)		EI (1-5)	
	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.
Howard	0.40	0.91	0.48	0.35	0.87	0.48	0.86	0.58	0.81	0.83
	F = 0.56		F = 0.40		**F = 0.62**		**F = 0.69**		**F = 0.82**	
AVS	0.37	0.89	0.42	0.29	0.84	0.47	0.84	0.59	0.77	0.80
	F = 0.52		F = 0.34		**F = 0.60**		**F = 0.69**		**F = 0.78**	

Table 3: Experimental results for entity assignment

	Next Sentences (Accuracy)				All Sentences (Accuracy)				Comp Ident.		
Data sets	Baseline1	Baseline2	ED (k-com)	ED (unk-com)	Baseline1	Baseline2	ED (k-com)	ED (unk-com)	Prec.	Recall	F
HowardForums	82.4%	83.3%	93.4%	90.3%	80.3%	82.1%	88.2%	86.7%	85.2%	84.2%	84.7%
AVSforrum	79.6%	80.9%	91.2%	89.6%	76.7%	77.9%	87.2%	85.0%	82.2%	84.9%	83.5%
Average	81.0%	82.1%	92.3%	89.9%	78.5%	80.0%	87.7%	85.9%	83.7%	84.6%	84.1%
Col#	1	2	3	4	5	6	7	8	9	10	11

[5] Khai phá câu so sánh

- Đánh giá của người dùng trên Internet đối với các sản phẩm:
  - 90% dưới dạng các đánh giá trực tiếp ("*the picture quality of Camera X is great*")
  - 10% số đánh giá của người dùng dưới dạng so sánh ("*the picture quality of Camera X is better than that of Camera Y*")

Phát biểu bài toán

- Nhiều câu so sánh không có từ so sánh trực tiếp, cảm xúc của cùng một từ phụ thuộc vào ngữ cảnh
- "*the battery life of Camera X is longer than Camera Y*"
- "*Program X's execution time is longer than Program Y*"
- Chọn câu làm ngữ cảnh dẫn đến bao gồm nhiều thông tin không liên quan

Phát biểu bài toán (2)

- Ngữ cảnh: thực thể được đánh giá + từ so sánh
- Làm thế nào xác định cảm xúc thể hiện bởi ngữ cảnh?
- Sử dụng tri thức từ bên ngoài (epinions.com) để xác định xu hướng cảm xúc từ ngữ cảnh
  - Epinions.com phân chia rõ ràng các mục bình luận tích cực và tiêu cực
  - Tìm xem ngữ cảnh hay xuất hiện trong các bình luận tích cực hay tiêu cực?

Phát biểu bài toán (3)

- Cho một quan hệ ứng với câu so sánh
  - <C (từ so sánh), F (đặc trưng), e1 (thực thể 1), e2 (thực thể 2), type (loại so sánh)>
- "*Camera X has longer battery life than Camera Y*"
  - <longer, battery life, Camera X, Camera Y, so sánh hơn>
- Xác định xem thực thể nào 'tốt' hơn

Pros và Cons

- **Pros**:
  - great photos <photo>
  - easy to use <use>
  - good manual <manual>
  - many options <option>
  - takes videos <video>

My SLR is on the shelf

by shortstop24, Aug 09 '03

Pros: Great photos, easy to use, good manual, many options, takes videos
Cons: Battery usage; included software could be improved; included 16MB is stingy.

I had never used a digital camera prior to purchasing the Canon A70. I have always used a SLR (Minol ...

Read the full review

Type 1 -er/-est

- Tính từ/trạng từ ngắn thêm hậu tố -er/est
- C thể hiện cảm xúc (better/best): Nếu cảm xúc tích cực, chọn e1, nếu cảm xúc tiêu cực, chọn e2
- C ko thể hiện cảm xúc, F thể hiện cảm xúc
  - "*Car X generates more noise than Car Y*"
  - C so sánh hơn + F tích cực → e1
  - C so sánh giảm + F tích cực → e2
  - C so sánh hơn + F tiêu cực → e2
  - C so sánh giảm + F tiêu cực → e1

Type 1 -er/-est (2)

- Cả C và F đều không có cảm xúc

$$OSA(F,C) = log \frac{Pr(F,C)Pr(C|F)}{Pr(F)Pr(C)}$$

  - Nếu $OSA_P(F,C) > OSA_N(F,C)$ → e1; ngược lại chọn e2

Type 1 -er/-est (3)

- Tính $OSA_P(F,C)$:
  - Đếm số lần C (và các từ đồng nghĩa) và F (và các từ đồng nghĩa) cùng xuất hiện trong Pros
  - Đếm số lần các từ trái nghĩa của C với F cùng xuất hiện trong Cons
  - Đếm số lần C và các từ trái nghĩa của F cùng xuất hiện trong Cons
- Sử dụng Wordnet để lấy từ đồng nghĩa và trái nghĩa
- Thực hiện tương tự với $OSA_N(F,C)$

Type 1 -er/-est (4)

- Nếu C thể hiện một đặc trưng
  - "*Camera X is smaller than Camera Y*"
  - Đếm số lần C xuất hiện ở Pros và Cons và chọn giá trị lớn hơn

Type 2 more/less + adj/adv

- Tính từ/trạng từ thể hiện cảm xúc
  - "*Car X has more beautiful interior than Car Y*"
- Tính từ/trạng từ không thể hiện cảm xúc
  - Tính từ/trạng từ mô tả đặc trưng
  - Tính từ/trạng từ không mô tả đặc trưng
- Phủ định:
  - "*Camera X's battery life is not longer than that of Camera Y*"

Đánh giá kết quả

- Dữ liệu gồm các bình luận đánh giá về camera, DVD players, MP3 players, Intel vs AMD, Coke vs Pepsi, Microsoft vs Google; laptop, mobile phone

Data Sources	No. of Comparative Sentences
(Jindal and Liu 2006)	418
Reviews and forum posts	419
Total	837

Đánh giá kết quả (2)

- Baseline-84%: Luôn lấy thực thể đầu tiên

	EntityS1 Preferred			**EntityS2 Preferred**		
	Prec.	*Rec.*	*F*	*Prec.*	*Rec.*	*F*
PCS (OSA)	0.967	0.966	0.966	0.822	0.828	0.825
PCS: No Pros & Cons	0.925	0.980	0.952	0.848	0.582	0.690
PCS (PMI)	0.967	0.961	0.964	0.804	0.828	0.816

	EntityS1 Preferred			**EntityS2 Preferred**		
	Prec.	*Rec.*	*F*	*Prec.*	*Rec.*	*F*
PCS (OSA)	0.896	0.877	0.886	0.696	0.736	0.716
PCS: No Pros & Cons	0.722	1.000	0.839	0.000	0.000	0.000
PCS (PMI)	0.894	0.855	0.874	0.661	0.736	0.696

ĐẠI HỌC BÁCH KHOA

25 YEARS ANNIVERSARY SOICT

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

Thank you for your attentions!

soict.hust.edu.vn/ fb.com/groups/soict