Nghĩa từ vựng và phân giải nhập nhằng từ

Từ đồng âm

- Từ đồng âm (Homonymy): là những từ trùng nhau về hình thức ngữ **âm** nhưng khác nhau về **nghĩa**
  - Từ đồng âm, đồng tự (Homograph) : các từ với cùng cách viết nhưng có nghĩa khác nhau.

    Ví dụ:
    - dove - dive into water, white bird
    - saw
  - Từ đồng âm, không đồng tự (Homophone): các từ có cách viết khác nhau nhưng có cùng âm.

    Ví dụ:
    - see, sea; meat, meet

Phân loại từ đồng âm tiếng Việt

- Đồng âm từ với từ, gồm:
  - Đồng âm từ vựng: Tất cả các từ đều thuộc cùng một từ loại. Ví dụ:
    - *đường*$_1$ (đắp đường) - *đường*$_2$ (đường phèn).
    - *đường kính*$_1$ (đường để ăn) - *đường kính*$_2$ (…của đường tròn).
    - *cất*$_1$ (cất vó) - *cất*$_2$ (cất tiền vào tủ) - *cất*$_3$ (cất hàng) - *cất*$_4$ (cất rượu)
  - Đồng âm từ vựng-ngữ pháp: Các từ trong nhóm đồng âm với nhau chỉ khác nhau về từ loại. Ví dụ:
    - *chỉ*$_1$ (cuộn chỉ) - *chỉ*$_2$ (chỉ tay năm ngón) - *chỉ*$_3$ (chỉ còn có dăm đồng).
    - *câu*$_1$ (nói vài câu) - *câu*$_2$ (rau câu) - *câu*$_3$ (chim câu) - *câu*$_4$ (câu cá)
- Đồng âm từ với tiếng: các đơn vị khác nhau về cấp độ; kích thước ngữ âm của chúng đều không vượt quá một tiếng. Ví dụ:
  - Con trai Văn *Cốc* lên dốc bắn cò, đứng lăm *le* cười khanh *khách*. Con gái Bát *Tràng* bán hàng thịt ếch ngồi châu *chẫu* nói ương *ương*.

Từ đa nghĩa, đồng nghĩa

- Từ đa nghĩa (Polysemy): một từ có thể có nhiều nghĩa mà cú pháp chỉ giúp phân biệt nghĩa đ/v các từ loại khác nhau của 1 từ nhập nhằng
  - *chỉ*$_1$ (cuộn chỉ) - *chỉ*$_2$ (chỉ tay năm ngón) - *chỉ*$_3$ (chỉ còn có dăm đồng).
  - “conduct” (noun or verb)
    - John’s conduct in class is unacceptable.
    - John will conduct the orchestra on Thursday.
- Đồng nghĩa (Synonymy): là những từ tương đồng với nhau về nghĩa, khác nhau về âm thanh. Ví dụ
  - cố, gắng
  - car, automobile

Nghĩa từ vựng

- Nghĩa của 1 từ là gì?
  - Homonyms (các nghĩa khác nhau)
    - bank: financial institution
    - bank: sloping land next to a river
  - Polysemes (các nghĩa có liên quan/gần nhau)
    - bank: financial institution as corporation
    - bank: a building housing such an institution
- Các nguồn ngữ liệu đ/v nghĩa từ:
  - Dictionaries (thesaurus)
  - Lexical databases

Nghĩa từ vựng

- Ngữ nghĩa nghiên cứu ý nghĩa của các phát biểu dạng ngôn ngữ
- Nghĩa từ vựng (Lexical semantics) nghiên cứu:
  - quan hệ từ vựng: sự liên hệ về mặt ngữ nghĩa giữa các từ
  - ràng buộc về lựa chọn: cấu trúc liên hệ ngữ nghĩa bên trong của từng từ
  - bao gồm lý thuyết về:
    - phân loại và phân rã nghĩa của từ
    - sự giống và khác trong cấu trúc từ vựng – ngữ nghĩa giữa các ngôn ngữ
    - quan hệ nghĩa của từ với cú pháp và ngữ nghĩa của câu.

Các ứng dụng

- Dịch máy
- Tóm tắt văn bản
- Phân loại văn bản
- Phân tích quan điểm
- Quảng cáo hướng ngữ cảnh
- Đối sánh văn bản
- Máy tìm kiếm
- Hệ thống hội thoại (dialogue system)
- Hệ thống hỏi đáp (question answering)
- …

Ràng buộc về lựa chọn

- Có rất nhiều từ đòi hỏi các bổ nghĩa (thường là các Động từ- các vị từ). Các bổ nghĩa này thường là các Danh từ và phải thỏa mãn các ràng buộc về lựa chọn.
- Ví dụ:
  - read (human subject, textual object)
  - eat (animate subject)
  - kill (animate object)
- Sử dụng vị từ để phân giải nhập nhằng ?
  - Một kiểu thông tin ngữ cảnh là thông tin về kiểu các bổ nghĩa mà 1 từ nhập nhằng yêu cầu.
  - Các vị từ khác nhau ứng với các nghĩa khác nhau
    - wash the **dishes** (theme : washable-thing)
    - serve vegetarian **dishe**s (theme : food-type)
  - Kiểu các bổ nghĩa cũng có thể giải quyết nhập nhằng cho vị từ

Đánh giá về các ràng buộc

- Yêu cầu liệt kê đầy đủ trong dạng máy có thể đọc được:
  - Cấu trúc bổ nghĩa của các Động từ.
  - Các ràng buộc về lựa chọn của các bổ nghĩa.
  - Mô tả các đặc tính của các từ đáp ứng được tiêu chí của ràng buộc về lựa chọn.
    - E.g. This flight serves the “**region”** between Mumbai and Delhi
    - How do you decide if “region” is compatible with “sector”
  - Sử dụng Từ điển đồng nghĩa hay Wordnet:
    - gồm từ đồng nghĩa (Synonyms) và trái nghĩa (Antonyms)
    - Từ lớp cha và từ lớp con
- Đô chính xác:
  - 44% on Brown corpus.

Đánh giá về các ràng buộc

- Các danh từ riêng (tên riêng) trong ngữ cảnh của 1 từ nhập nhằng có thể xem như dấu hiện xử lý nhập nhằng rất mạnh.

  E.g. **“Sachin Tendulkar”** will be a strong indicator of the category **“sports”.**

  Sachin Tendulkar plays **cricket.**

- Các danh từ riêng không xuất hiện trong thesaurus hay Wordnet. Từ đó cách tiếp cận này không khai thác được các dấu hiệu mạnh của các danh từ riêng.
- Độ chính xác:
  - 50% khi được test trên 10 từ có nhiều nghĩa trong tiếng Anh.

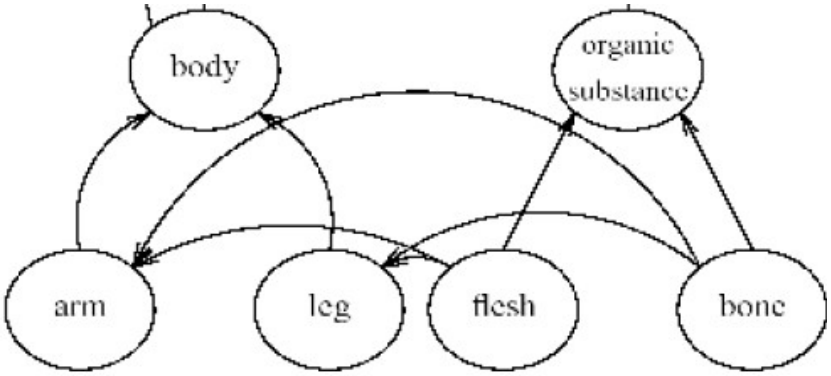
Đánh giá về các ràng buộc

- Ưu điểm
  - Một tiếp cận không phân tích cú pháp.
  - Cài đặt đơn giản.
  - Không yêu cầu 1 bộ dữ liệu đ/v từ nhập nhằng.
- Nhược điểm
  - Có thể gặp đối sánh thưa*:* khả năng bao trùm từ là rất ít.
  - Không sử dụng được với các trường hợp không liệt kê trong máy.
  - Các danh từ riêng (tên riêng) trong ngữ cảnh của 1 từ nhập nhằng có thể xem như dấu hiện xử lý nhập nhằng rất mạnh nhưng các danh từ riêng không xuất hiện trong thesaurus. Từ đó cách tiếp cận này không khai thác được các dấu hiệu mạnh của các danh từ riêng.

Đánh giá về các ràng buộc

- Vấn đề:
  - Đôi khi ràng buộc lựa chọn không đủ chặt (khi 1 từ có nhiều nghĩa)
  - Đôi khi ràng buộc quá chặt – khi vị từ sử dụng phép ẩn dụ. Vd, I’ll eat my hat!

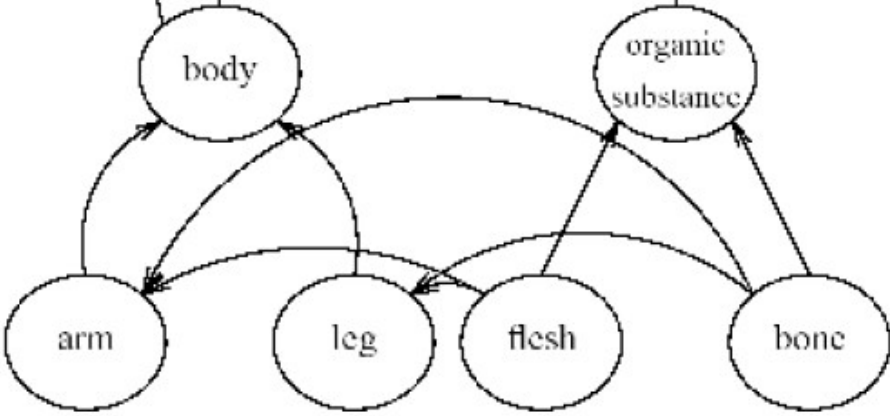
WordNet: Giới thiệu



CSDL từ vựng

- Xây dựng một mạng khổng lồ các từ vựng và quan hệ giữa các từ vựng
- Wordnet tiếng Anh
  - 4 lớp: danh từ, động từ, tính từ, trạng từ
  - Danh từ: 120,000; Động từ: 22,000; Tính từ: 30,000;
  - Trạng từ: 6,000

WordNet: Giới thiệu



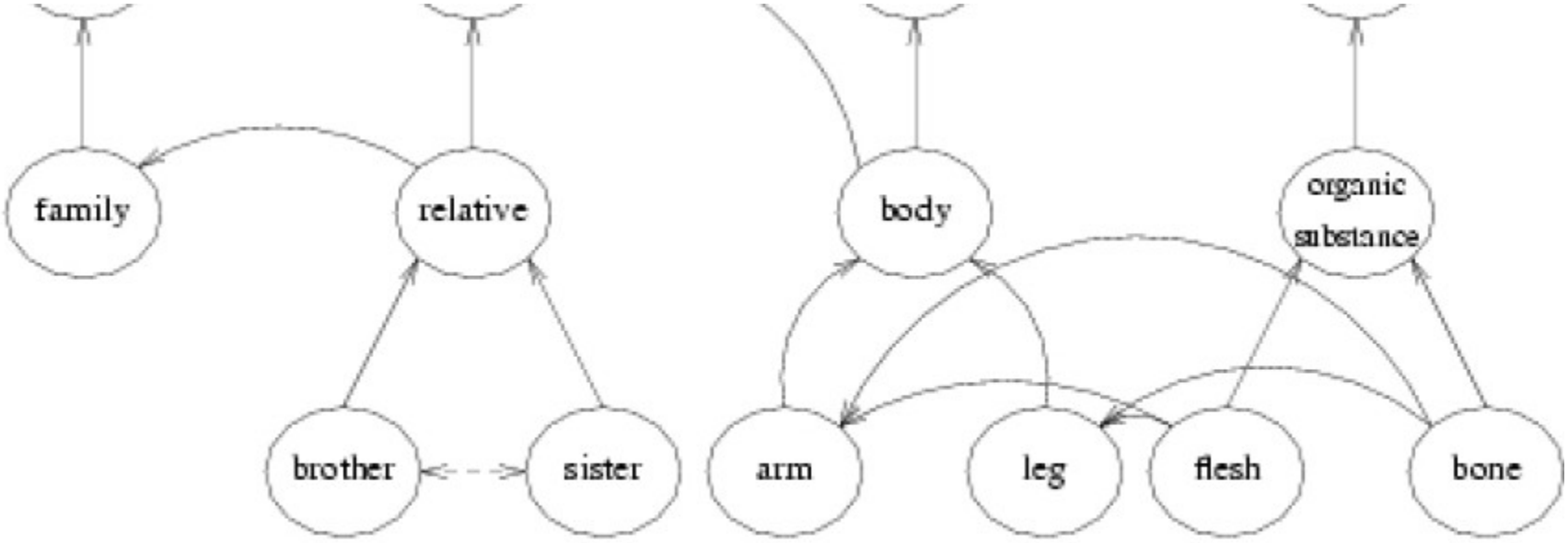
- CSDL từ vựng
  - Wordnet cho các ngôn ngữ khác [www.globalwordnet.org]
    - Có wordnet cho các ngôn ngữ: Tây Ban Nha, Tiệp, Hà Lan, Pháp, Đức, Ý, Bồ Đào Nha, Thụy Điển, Basque, Estonian
    - Wordnets đang được làm cho các tiếng: Bulgary, Đan mạch, Hy lạp, Hebrew, Hindi, Cannada, Latvian, Moldavy, Romany, Nga, Slovenian, Tamil, Thái lan, Thổ Nhĩ Kỳ, Ireland, Nauy, Ba tư, Iran

Tập từ đồng nghĩa
Synonym Sets - Synsets

- Từ có nhập nhằng
- Các nút trong Wordnet biểu diễn tập từ đồng nghĩa “synonym sets”, hoặc *synsets.* Ví dụ:
  - Fool: 1 người dễ bị lợi dụng
  - {chump, fish, fool, gull, mark, patsy, fall guy, sucker, schlemiel, shlemiel, soft touch, mug}
  - Synset = tập khái niệm

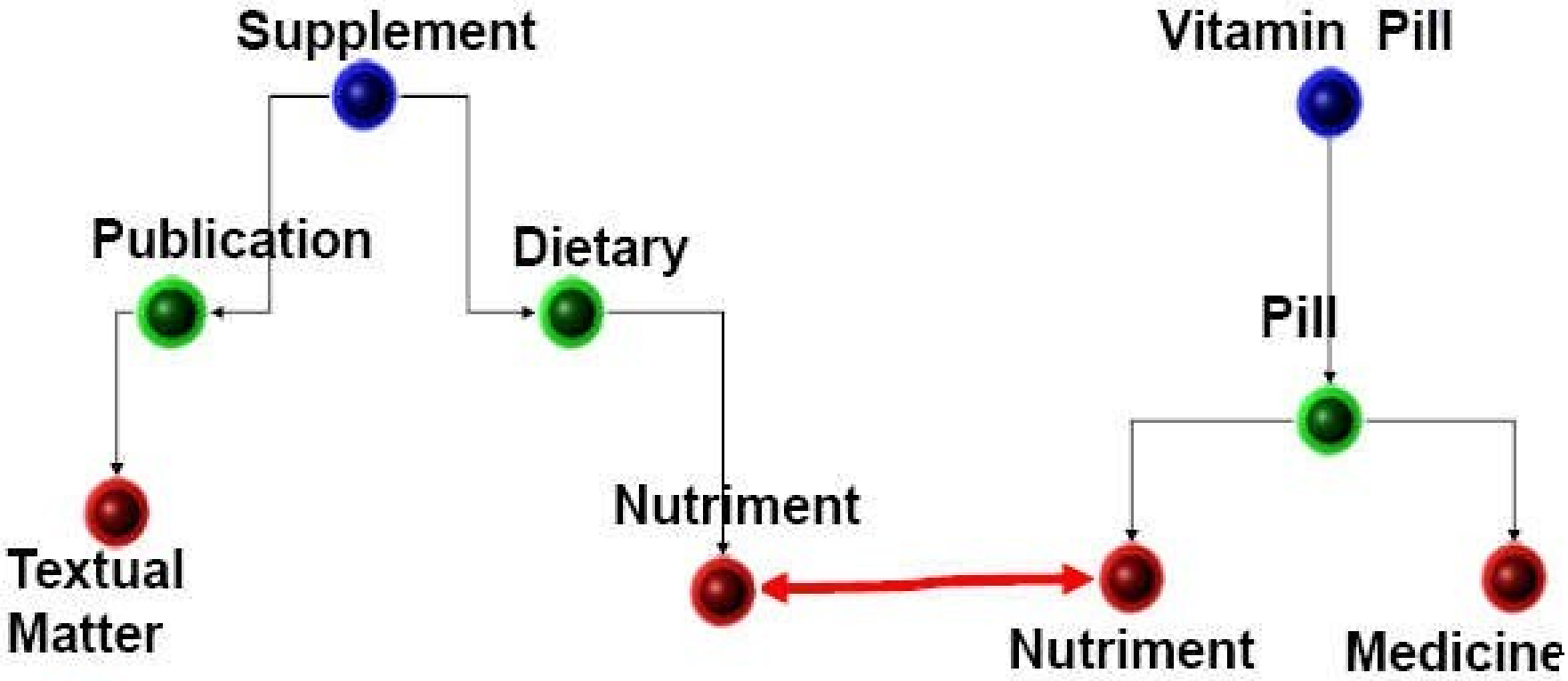
Các quan hệ khác trong WordNet

- Các từ nối theo chiều dọc biểu diễn quan hệ rộng (holonymy) - hẹp (hypernymy), theo chiều ngang biểu diễn quan hệ bộ phận meronymy (part_of) và holonymy (has_part) .
- Mỗi nghĩa của từ được biểu diễn bằng 1 số synset

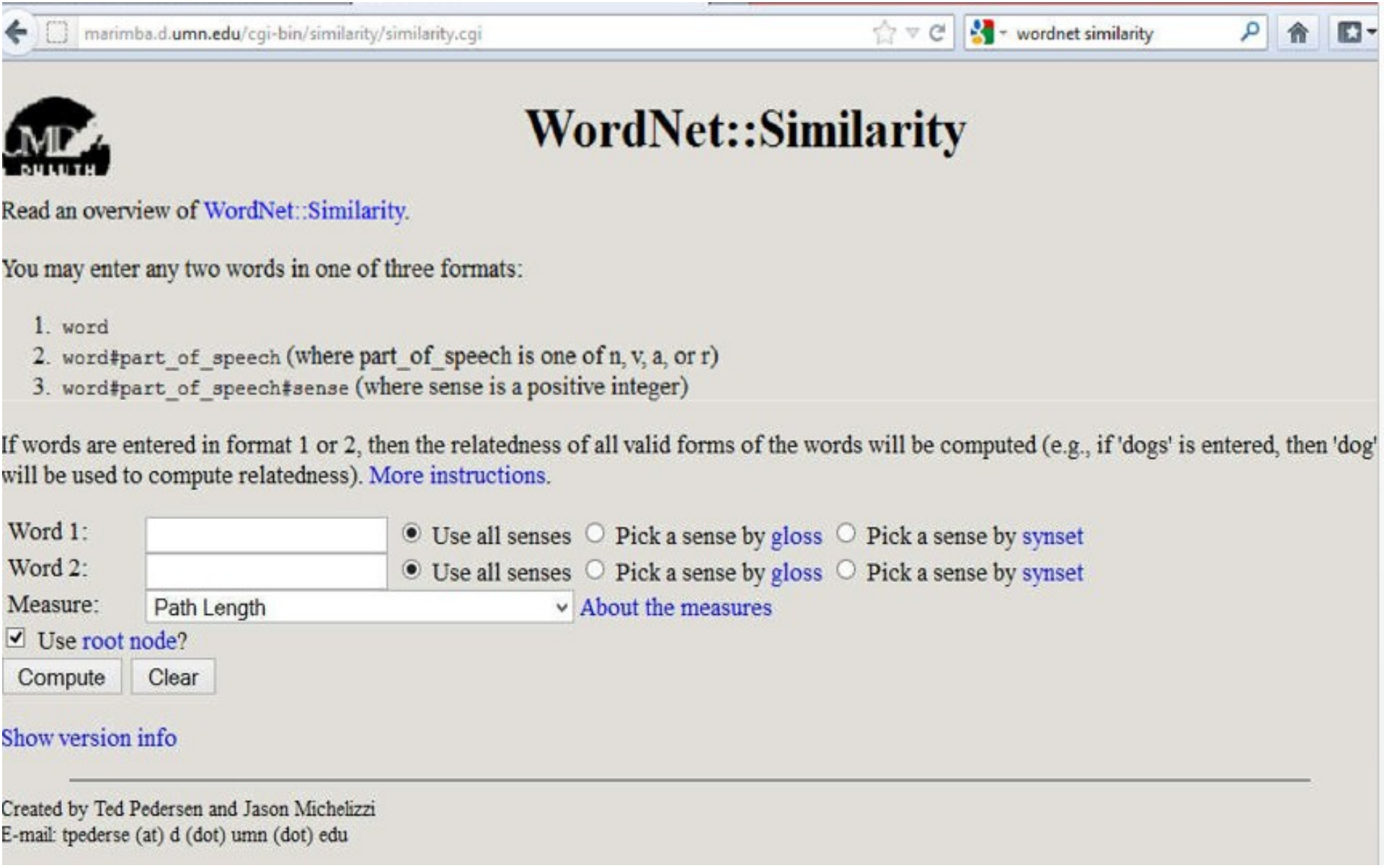


16

Phân giải nhập nhằng sử dụng quan hệ từ vựng

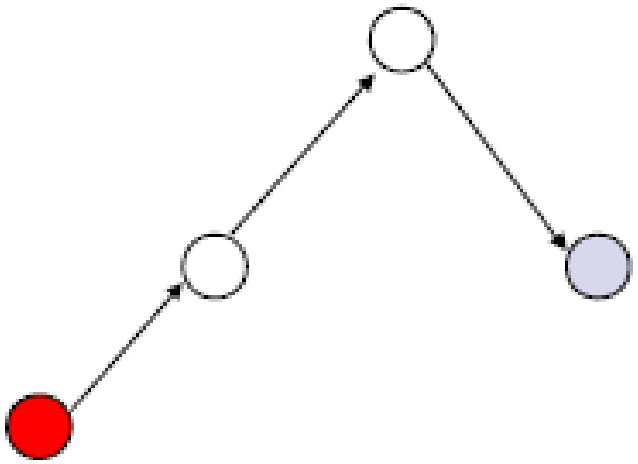


WordNet Similarity Metrics:
http://marimba.d.umn.edu/cgi-bin/similarity/similarity.cgi


marimba.d.umn.edu/cgi-bin/similarity/similarity.cgi
wordnet similarity
WordNet::Similarity
Read an overview of WordNet::Similarity.
You may enter any two words in one of three formats:
1. word
2. word#part_of_speech (where part_of_speech is one of n, v, a, or r)
3. word#part_of_speech#sense (where sense is a positive integer)
If words are entered in format 1 or 2, then the relatedness of all valid forms of the words will be computed (e.g., if 'dogs' is entered, then 'dog' will be used to compute relatedness). More instructions.
Word 1:
Use all senses
Pick a sense by gloss
Pick a sense by synset
Word 2:
Use all senses
Pick a sense by gloss
Pick a sense by synset
Measure:
Path Length
About the measures
Use root node?
Compute
Clear
Show version info
Created by Ted Pedersen and Jason Michelizzi
E-mail: tpederse (at) d (dot) umn (dot) edu

Đo quan hệ từ vựng

- Đếm số cạnh/đỉnh trên đồ thị:
  - khoảng cách giữa 2 từ tỉ lệ nghịch với quan hệ ngữ nghĩa giữa chúng
  - Nếu giữa 2 từ có nhiều đường đi, chọn đường ngắn nhất



số cạnh = 3

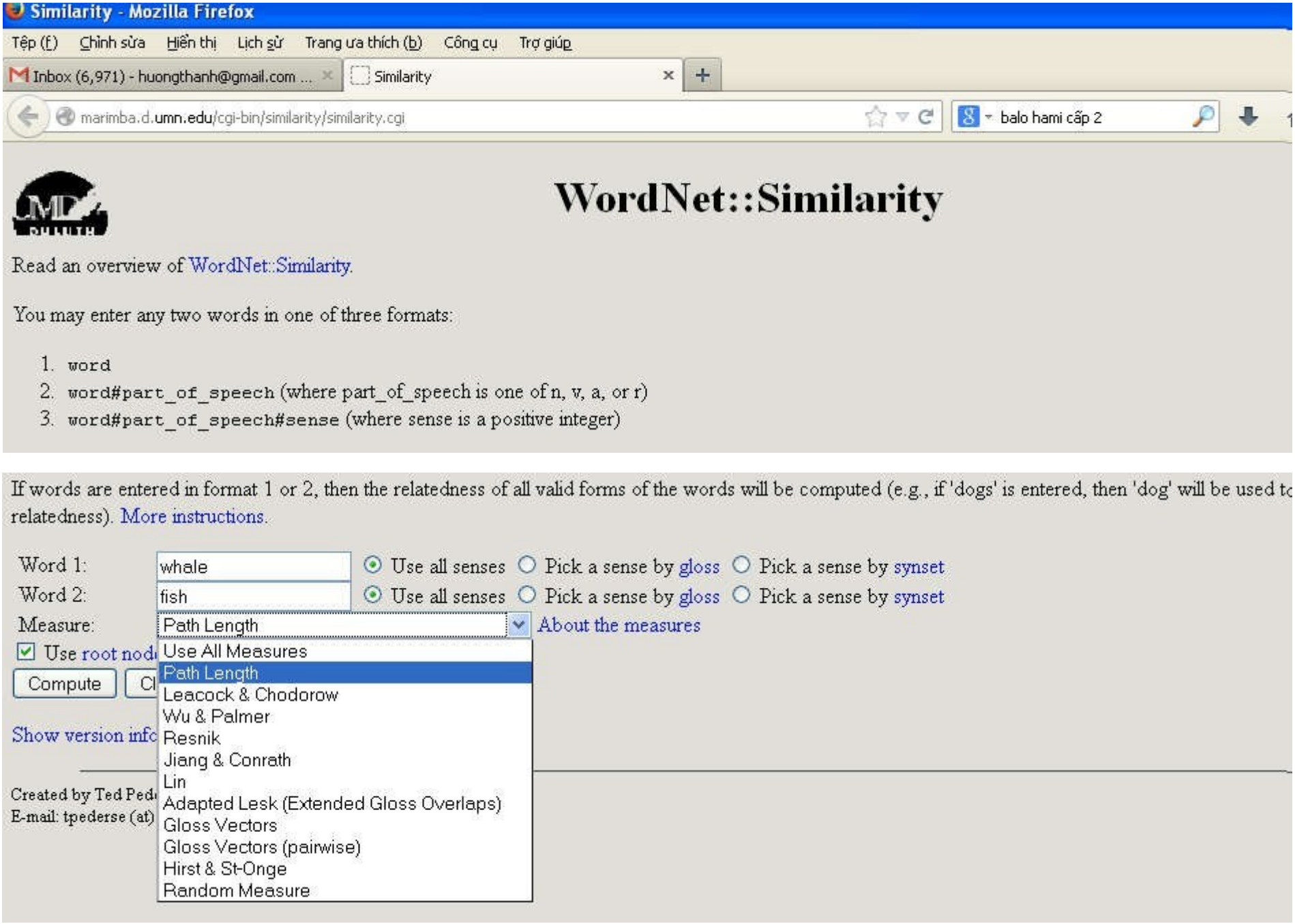
số nút = 4

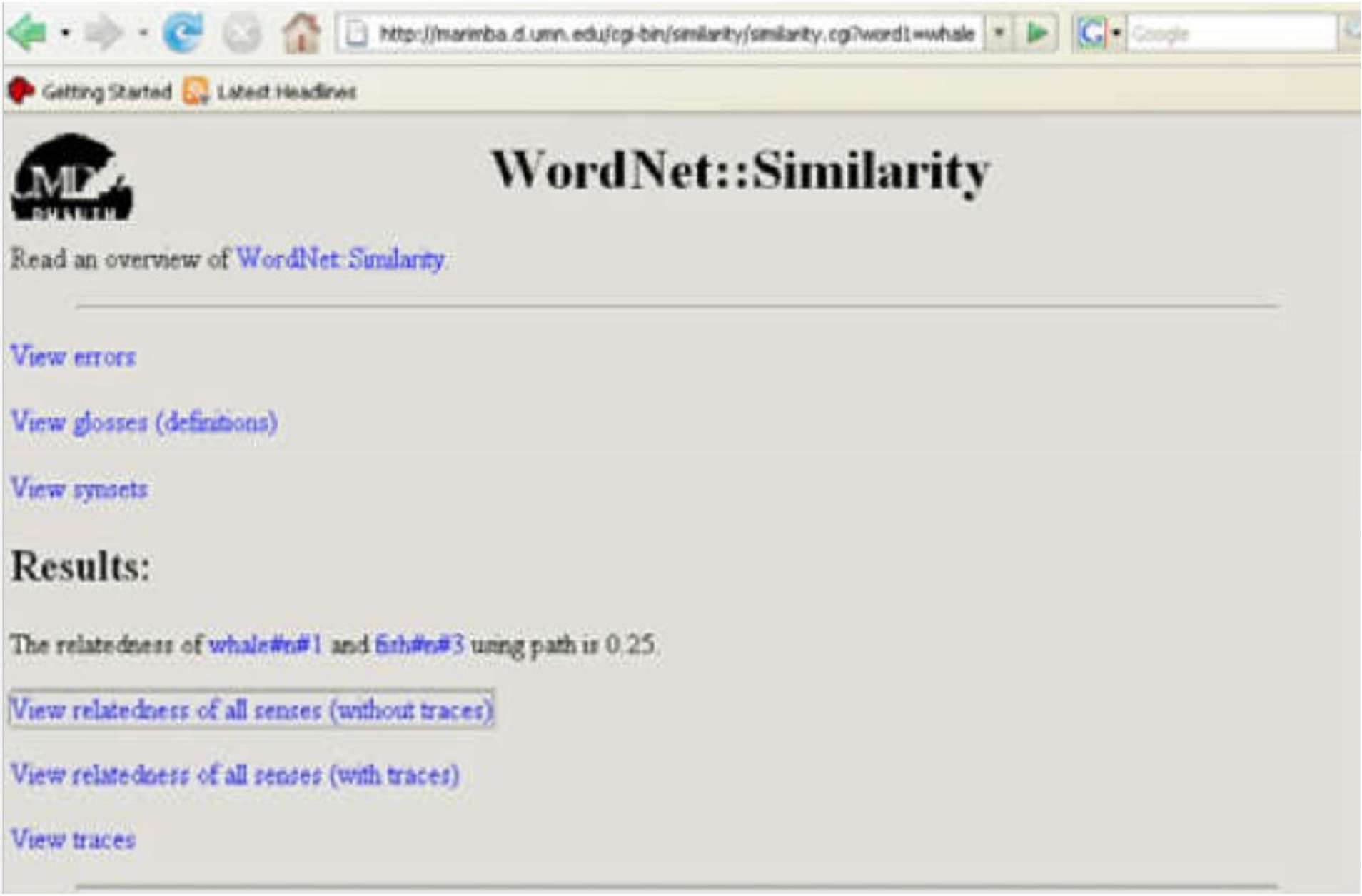
Cặp từ nào gần nhau hơn?

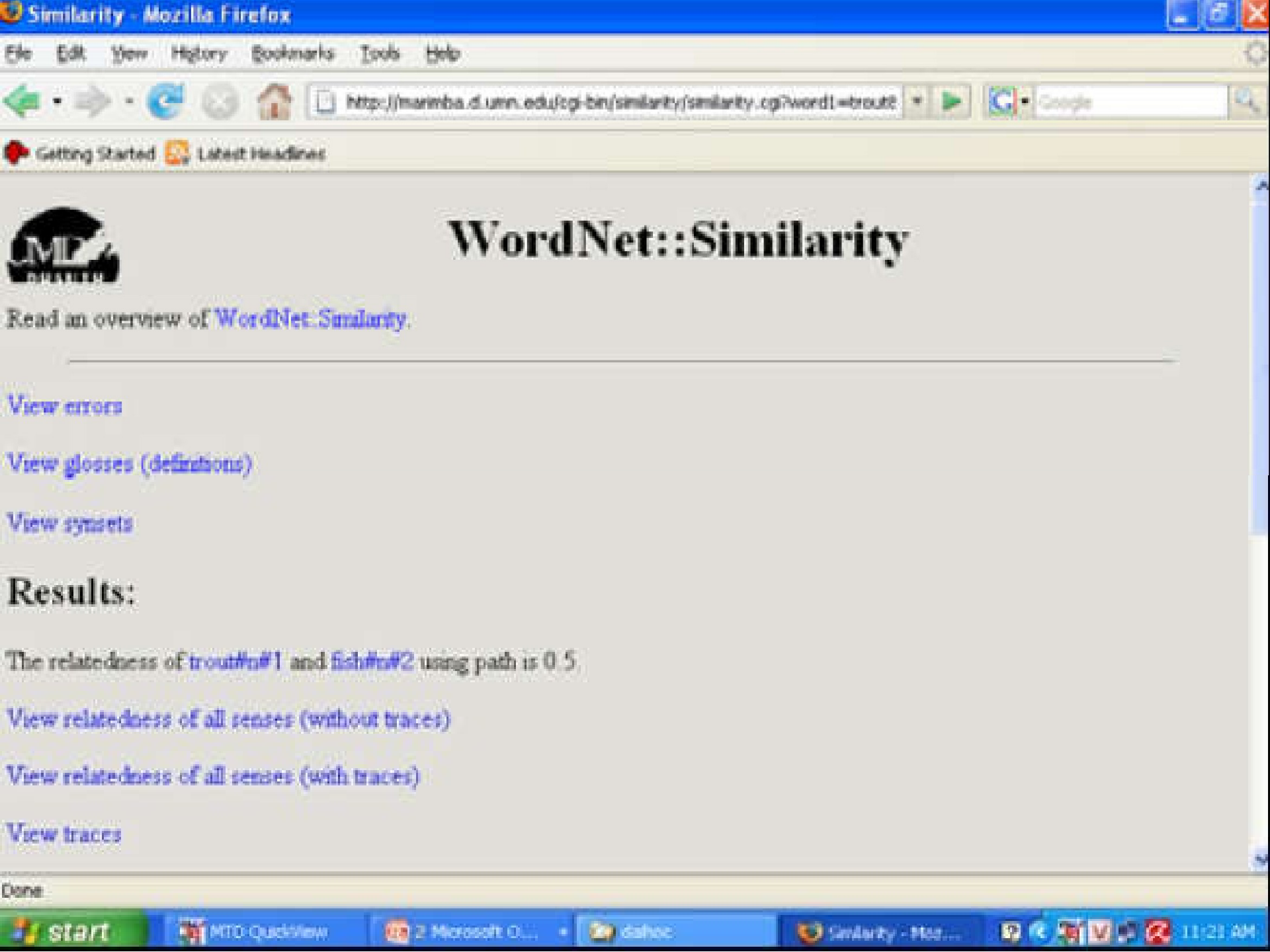
- cá heo và cá?
- cá và cá hồi?

WordNet Similarity Metrics:

http://marimba.d.umn.edu/cgi-bin/similarity/similarity.cgi

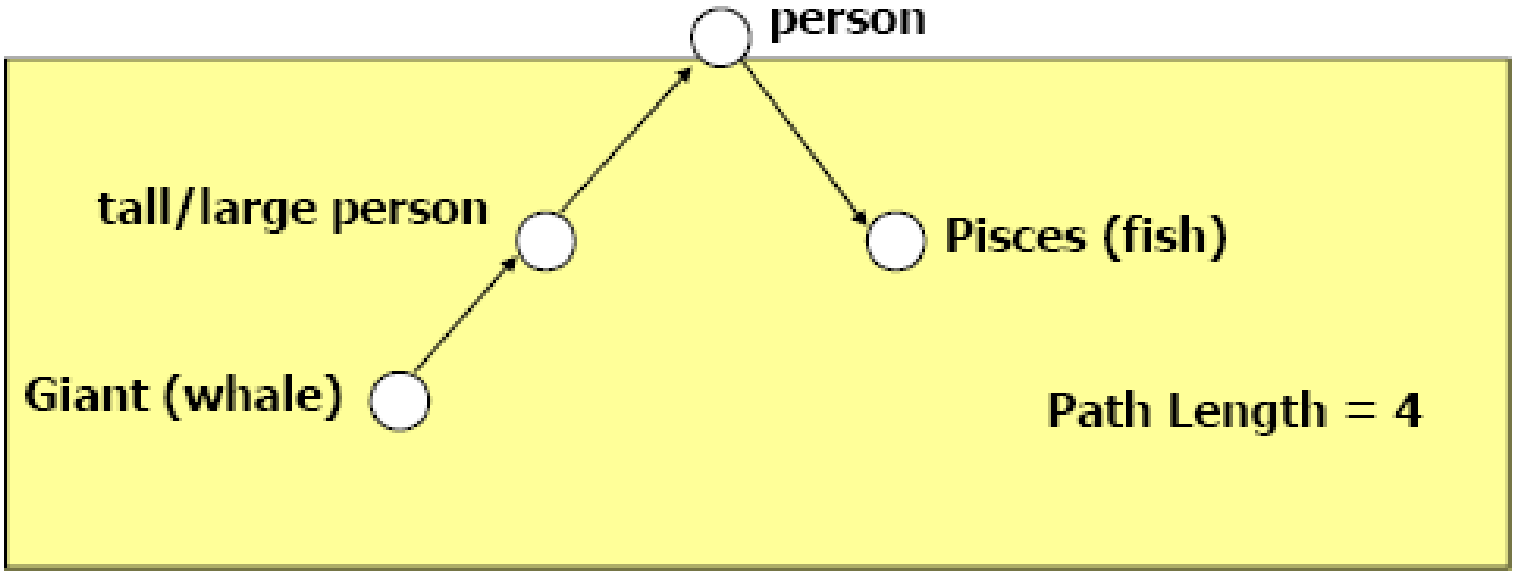

Similarity - Mozilla Firefox
Tệp (f) Chỉnh sửa Hiển thị Lịch sử Trang ưa thích (b) Công cụ Trợ giúp
Inbox (6,971) - huongthanh@gmail.com ...
Similarity
marimba.d.umn.edu/cgi-bin/similarity/similarity.cgi
balo hami cấp 2
WordNet::Similarity
Read an overview of WordNet::Similarity.
You may enter any two words in one of three formats:
1. word
2. word#part_of_speech (where part_of_speech is one of n, v, a, or r)
3. word#part_of_speech#sense (where sense is a positive integer)
If words are entered in format 1 or 2, then the relatedness of all valid forms of the words will be computed (e.g., if 'dogs' is entered, then 'dog' will be used to
relatedness). More instructions.
Word 1: whale
Use all senses
Pick a sense by gloss
Pick a sense by synset
Word 2: fish
Use all senses
Pick a sense by gloss
Pick a sense by synset
Measure: Path Length
About the measures
Use root nod
Use All Measures
Path Length
Leacock & Chodorow
Wu & Palmer
Resnik
Jiang & Conrath
Lin
Adapted Lesk (Extended Gloss Overlaps)
Gloss Vectors
Gloss Vectors (pairwise)
Hirst & St-Onge
Random Measure
Compute
Cl
Show version info
Created by Ted Ped
E-mail: tpederse (at)


http://maximba.d.umn.edu/cgi-bin/similarity/similarity.cgi?word1=whale
Getting Started
Latest Headlines
WordNet::Similarity
Read an overview of WordNet::Similarity
View errors
View glosses (definitions)
View synsets
Results:
The relatedness of whale#n#1 and fish#n#3 using path is 0.25
View relatedness of all senses (without traces)
View relatedness of all senses (with traces)
View traces

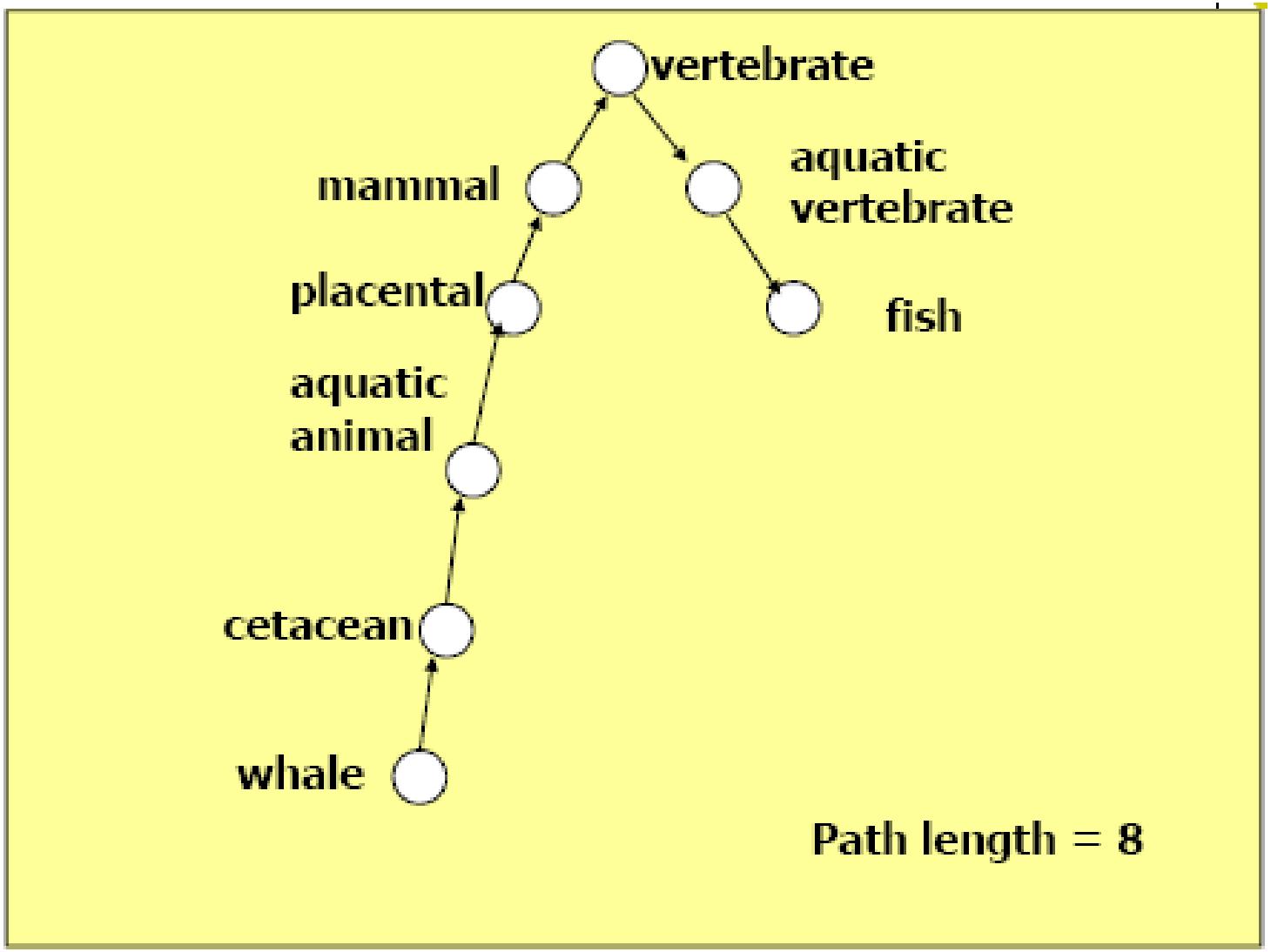

Similarity - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Getting Started Latest Headlines
WordNet::Similarity
Read an overview of WordNet::Similarity.
View errors
View glosses (definitions)
View synsets
Results:
The relatedness of trout#n#1 and fish#n#2 using path is 0.5
View relatedness of all senses (without traces)
View relatedness of all senses (with traces)
View traces
Done
start

Phân giải nhập nhằng và đếm cạnh

- whale#n#1
  - 1 người rất lớn (về kích thước hoặc phẩm chất)
- fish#n#3
  - (thiên văn học) người được sinh khi mặt trời ở vì sao Pisces



Phân giải nhập nhằng và đếm cạnh



Nhược điểm của WordNet trong tính quan hệ ngữ nghĩa

- Độ đo quan hệ ngữ nghĩa WordNet dựa trên các giả thiết sau:
  - Mọi cạnh trong đồ thị có độ dài bằng nhau
  - Các nhánh trong đồ thị có cùng độ đậm đặc
  - Tồn tại tất cả các quan hệ ngoại động từ

➢ không đáng tin cậy

Cách tiếp cận dựa trên từ điển

- Các từ điển điện tử (Lesk ‘86)
  - Cho biết ý nghĩa của các từ trong ngữ cảnh cụ thể nội dung (vd., I’ve often caught bass while out at sea)
  - So sánh sự chồng chéo của các định nghĩa về nghĩa của từ ($bass_2$: a type of fish that lives in the sea)
  - Chọn nghĩa trùng nhau nhiều nhất
- Hạn chế: đường dẫn đến từ ngắn mở rộng cho các từ liên quan

Cách tiếp cận học máy

- Học việc phân loại để gán từ với một trong các nghĩa của nó
  - Tích lũy tri thức từ tập ngữ liệu có hoặc không gán nhãn
  - Con người chỉ can thiệp vào tập ngữ liệu gán nhãn và lựa chọn tập đặc trưng sử dụng trong việc huấn luyện
- Vào: vectơ đặc trưng
  - đích (từ cần phân giải nhập nhằng)
  - nội dung (các đặc trưng có thể dùng để tiên đoán nghĩa đúng)
- Ra: các luật phân loại cho văn bản mới

Các đặc trưng sử dụng trong WSD

- Các thẻ POS của từ và các từ lân cận
- Các từ lân cận (có thể lấy gốc từ hoặc không)
- Dấu chấm, viết hoa, định dạng
- PTCP bộ phận để xác định vai trò ngữ pháp và quan hệ giữa chúng
- Các thông tin về đồng xuất hiện:
  - Từ và các từ lân cận của nó có thường đồng xuất hiện không
- Đồng xuất hiện của các từ láng giềng
  - Ví dụ: sea có thường xuyên xuất hiện với bass không

Ví dụ

- Tôi ăn cơm với cá.
  - DT ĐgT DT GT DT
  - (C (CN (ĐaT Tôi)) (VN (ĐgN (ĐgN (ĐgT ăn) (DT cơm)) (GN (GT
    với) (DT cá)))))
- Em bé chỉ thích ăn kẹo thôi.
  - DT TT TT ĐgT DT PT
  - (C (CN (DT Em bé)) (VN (TN (TN (TT chỉ) (TN (TT thích) (ĐgN (ĐgT ăn) (DT kẹo)))) (PT thôi))))
- Nó ăn nhiều hoa hồng quá.
  - ĐaT ĐgT TT DT TT
  - (C (CN (ĐaT Nó)) (VN (ĐgN (ĐgN (ĐgT ăn) (TT nhiều) (DT hoa hồng)) (TT quá))))
- Tôi tên là Hoa.

Các kiểu phân loại

- Naïve Bayes: Nghĩa tốt nhất là nghĩa có khả năng xảy ra nhất với 1 đầu vào cho trước

$$\hat{s} = \arg\max_{s \in S} p(s|V), \text{ hoặc } \arg\max_{s \in S} \frac{p(V|s)p(s)}{p(V)}$$

  - trong đó s là 1 trong các nghĩa và V là vector đầu vào của các đặc trưng
  - Chỉ có ít dữ liệu có thông tin vector kết hợp với nghĩa
  - Giả sử các đặc trưng là độc lập, p(V|s) là tích xác suất của các đặc trưng

$$p(V|s) = \prod_{j=1}^{n} p(v_j|s)$$

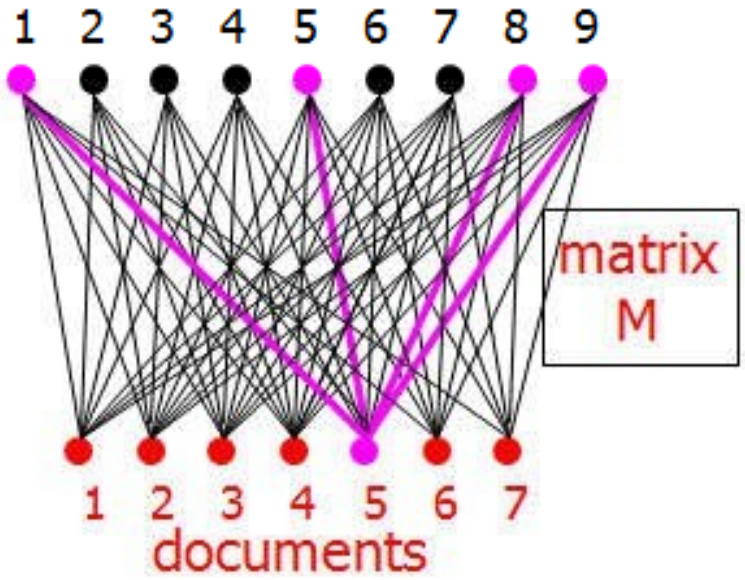
  - p(V) giống nhau với mọi ŝ (không ảnh hưởng đến xếp hạng cuối cùng)

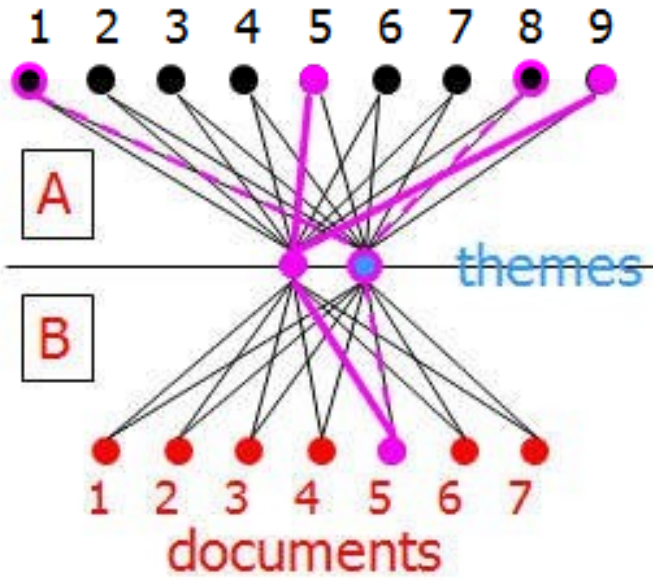
Các kiểu phân loại

- Naïve Bayes : Nghĩa tốt nhất là nghĩa có khả năng xảy ra nhất với 1 đầu vào cho trước
  - Khi đó
  - P(s) là xác suất tiền nghiệm của nghĩa s = xác suất của nghĩa s trong tập dữ liệu gán nhãn
  - P(v, s) = đếm số lần xuất hiện của v đi với s

Học máy xác định tập từ đồng nghĩa

- Phương pháp phân tích ngữ nghĩa tiềm ẩn:
  - SVD (Singular Value Decomposition)





Học máy xác định tập từ đồng nghĩa

- Phương pháp phân tích ngữ nghĩa tiềm ẩn:
  - LSA (Latent Semantic Analysis)

$$
\begin{array}{cccccc}
 & X & & U & \Sigma & V^T \\
 & (\mathbf{d}_j) & & & & (\hat{\mathbf{d}}_j) \\
 & \downarrow & & & & \downarrow \\
(\mathbf{t}_i^T) \rightarrow & \begin{bmatrix} x_{1,1} & \cdots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \cdots & x_{m,n} \end{bmatrix} & = & (\hat{\mathbf{t}}_i^T) \rightarrow \begin{bmatrix} \begin{bmatrix} \\ \mathbf{u}_1 \\ \\ \end{bmatrix} & \cdots & \begin{bmatrix} \\ \mathbf{u}_l \\ \\ \end{bmatrix} \end{bmatrix} & \cdot \begin{bmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_l \end{bmatrix} & \cdot \begin{bmatrix} [& \mathbf{v}_1 &] \\ & \vdots & \\ [& \mathbf{v}_l &] \end{bmatrix}
\end{array}
$$

Học máy xác định tập từ đồng nghĩa

- Phương pháp phân tích ngữ nghĩa tiềm ẩn:
  - LDA (Latent Dirichlet Allocation)

α is the parameter of the Dirichlet prior on the per-document topic distributions,

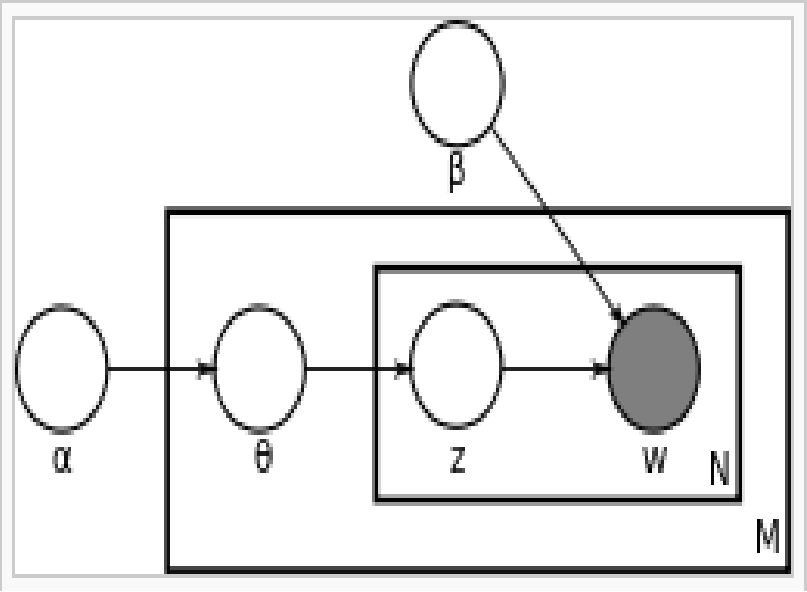
β is the parameter of the Dirichlet prior on the per-topic word distribution,

θ_i is the topic distribution for document i,

φ_k is the word distribution for topic k,

z_{ij} is the topic for the jth word in document i, and

w_{ij} is the specific word.



Ví dụ đầu ra của LDA

Topic0:	Topic1:	Topic2:	Topic3:	Topic4:	Topic5:	Topic6:	Topic7:	Topic8:	Topic9:
cstag	cstag	cstag	cstag	cstag	cstag	cstag	cstag	cstag	cstag
credit	agenttag	agenttag	agenttag	agenttag	agenttag	agenttag	agenttag	agenttag	agenttag
agenttag	credit	credit	credit	credit	interest	credit	interest	credit	credit
interest	interest	interest	interest	interest	credit	interest	credit	interest	interest
pay	pay	time	pay	payment	pay	pay	pay	pay	time
time	low	payment	time	pay	time	payment	payment	time	low
payment	rate	pay	low	time	low	time	time	low	payment
low	time	rate	payment	low	payment	low	rate	payment	rate
rate	payment	low	rate	rate	rate	use	low	use	pay
use	charg	use	use	use	use	rate	use	rate	use

Học máy xác định từ đồng nghĩa

- **Word embedding:** các kỹ thuật học mô hình ngôn ngữ và học đặc trưng với mỗi từ/cụm từ được biểu diễn bởi 1 vector các số thực trong không gian từ vựng
- Nhắc lại các phương pháp biểu diễn trước thời Word embedding:
  - One-hot encoding
  - Co-occurrence matrix

One-hot encoding

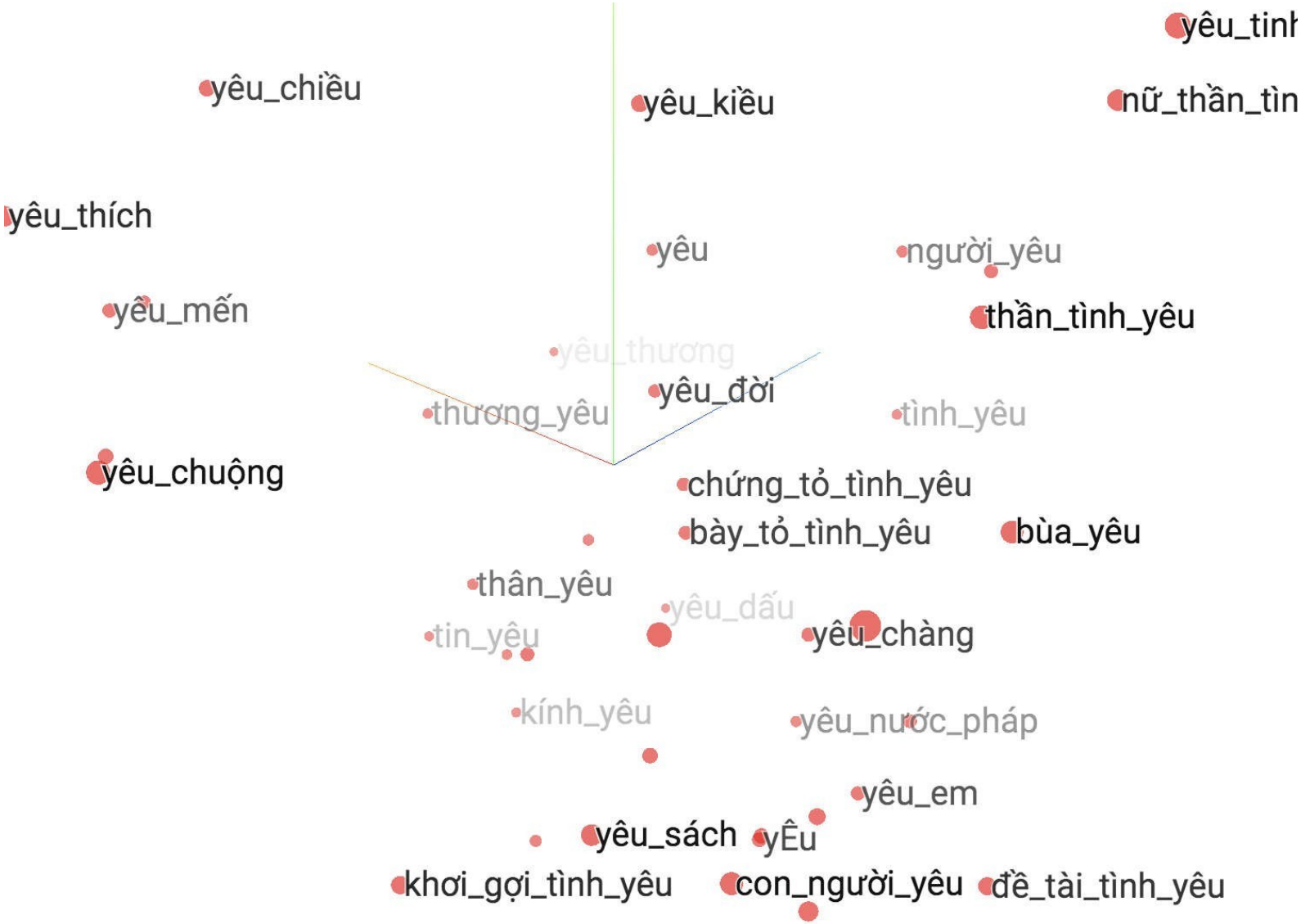
- Tập dữ liệu:
  - Tôi đang đi tìm một_nửa của mình
  - Tôi đã ăn một_nửa quả táo
  - Tôi đã đi tìm một_nửa quả táo
- Từ điển V= { tôi_1, đang_2, đi_3, tìm_4, một−nửa_5, của_6, mình_7, đã_8, ăn_9, quả_10, táo_11}
- Biểu diễn từ

  Tôi = [1 0 0 0 0 0 0 0 0 0 0]
  đang = [0 1 0 0 0 0 0 0 0 0 0]
  …
  mình = [0 0 0 0 0 0 1 0 0 0 0]
  táo = [0 0 0 0 0 0 0 0 0 0 1]
- Nhược: tốn tài nguyên, không thể hiện được liên hệ ngữ nghĩa giữa các từ

Tham khảo: https://forum.machinelearningcoban.com/t/hoc-bieu-dien-ngon-ngu-cho-may-tinh/299

Co-occurrence matrix

- “Bạn sẽ hiểu một từ qua các từ đi cùng với nó”



Tham khảo: https://forum.machinelearningcoban.com/t/hoc-bieu-dien-ngon-ngu-cho-may-tinh/299

Co-occurrence matrix

Tôi đang đi tìm một_nửa của mình
Tôi đã ăn một_nửa quả táo
Tôi đã đi tìm một_nửa quả táo

- Mức **văn bản** cho thông tin chung về các chủ đề hướng tới các phương pháp LSA
- Mức **cửa sổ từ** cho thông tin về cả chức năng **cú pháp** của từ và **ngữ nghĩa**.

Counts	tôi	đang	đi	tìm	một_nửa	của	mình	đã	ăn	quả	táo
tôi	0	1	0	0	0	0	0	2	0	0	0
đang	1	0	1	0	0	0	0	0	0	0	0
đi	0	1	0	2	0	0	0	1	0	0	0
tìm	0	0	2	0	2	0	0	0	0	0	0
một_nửa	0	0	0	2	0	1	0	0	1	2	0
của	0	0	0	0	1	0	1	0	0	0	0
mình	0	0	0	0	0	1	0	0	0	0	0
đã	2	0	1	0	0	0	0	0	1	0	0
ăn	0	0	0	0	1	0	0	1	0	0	0
quả	0	0	0	0	2	0	0	0	0	0	2
táo	0	0	0	0	0	0	0	0	0	2	0

Tham khảo: https://forum.machinelearningcoban.com/t/hoc-bieu-dien-ngon-ngu-cho-may-tinh/299

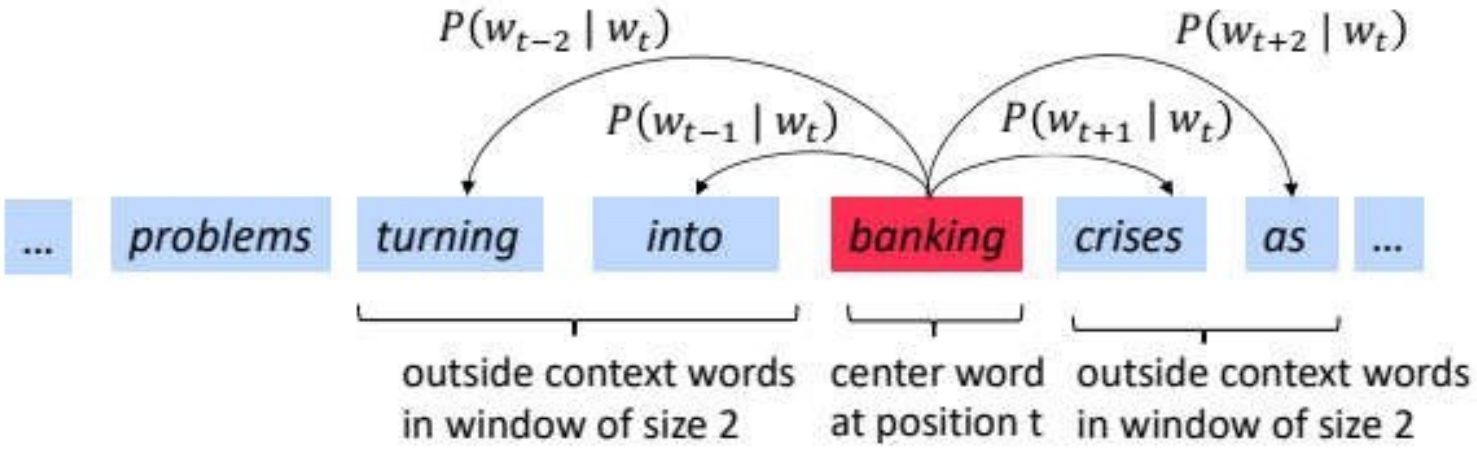
Co-occurrence matrix

- Ghi nhận được thông tin đồng xuất hiện của các từ trong dữ liệu học
- Vấn đề :
  - Chiều của vector tăng theo kích thước từ điển.
  - Cần không gian nhớ lớn để lưu thông tin.
  - Các mô hình phân loại sau đó dựa trên cách biểu diễn này sẽ gặp phải vấn đề biểu diễn thưa (sparsity issues).
- Giải quyết: Singular Value Decomposition

Tham khảo: https://forum.machinelearningcoban.com/t/hoc-bieu-dien-ngon-ngu-cho-may-tinh/299

Word embedding

- Thay vì lưu thông tin xuất hiện của các từ bằng cách đếm trực tiếp như **ma trận đồng xuất hiện**, word2vec học để đoán từ lân cận của tất cả các từ.
- Phương pháp:
  - Đoán các từ lân cận trong cửa số m của mỗi từ:
    - Với mỗi từ t = 1 … T, đoán các từ trong cửa sổ bán kính m của tất cả các từ



Hàm mục tiêu

For each position $t = 1, \ldots, T$, predict context words within a window of fixed size m, given center word w_j.

$$\text{Likelihood} = L(\theta) = \prod_{t=1}^{T} \prod_{\substack{-m \le j \le m \\ j \neq 0}} P(w_{t+j} \mid w_t; \theta)$$

θ is all variables to be optimized

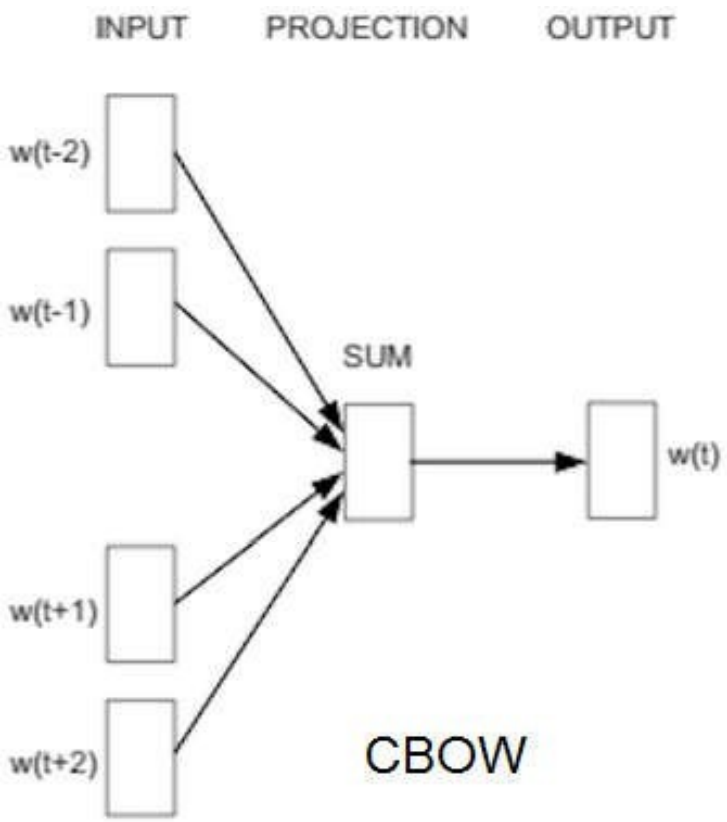
sometimes called *cost* or *loss* function

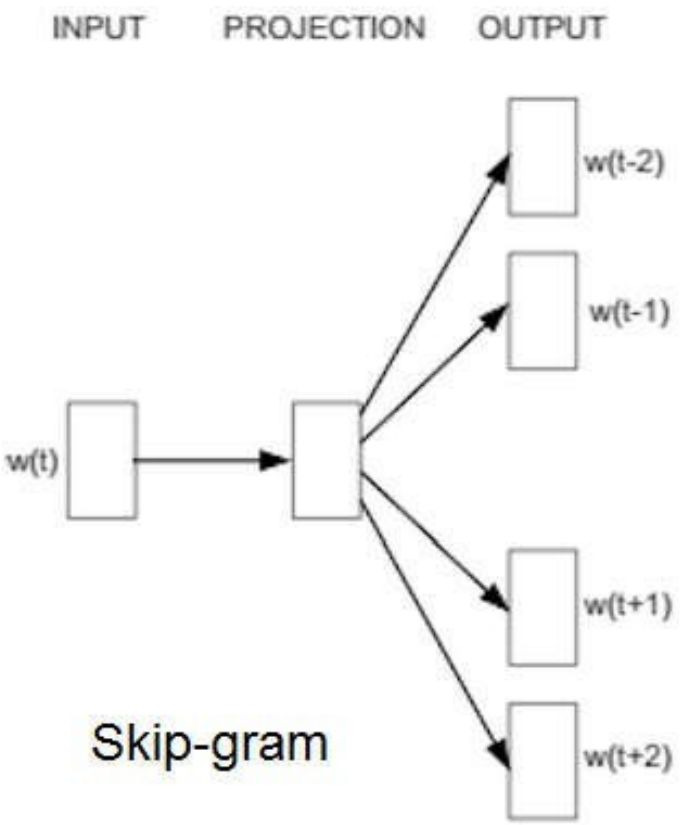
The objective function $J(\theta)$ is the (average) negative log likelihood:

$$J(\theta) = -\frac{1}{T} \log L(\theta) = -\frac{1}{T} \sum_{t=1}^{T} \sum_{\substack{-m \le j \le m \\ j \neq 0}} \log P(w_{t+j} \mid w_t; \theta)$$

Hàm giá (loss/cost function)

Word embedding





CBoW:

- Cho các từ ngữ cảnh
- Đoán xác suất của một từ đích

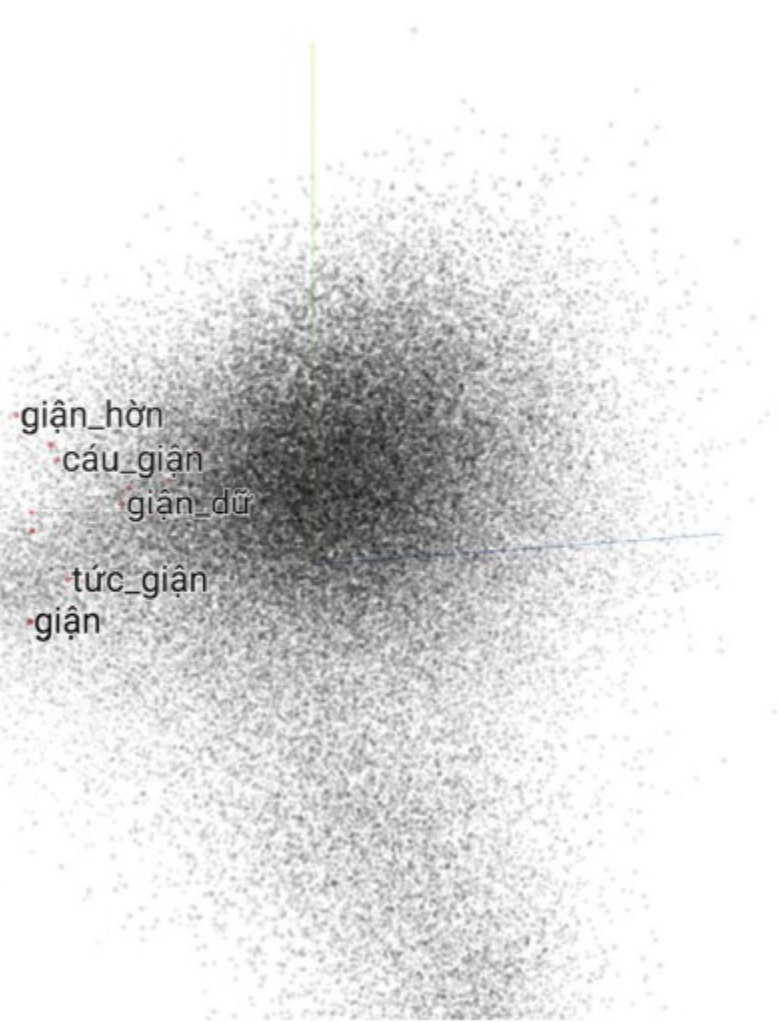
Skip-gram:

- Cho từ đích
- Đoán xác suất của các từ ngữ cảnh

Word embedding

- Các phiên bản:
  - Gensim: đầu vào là các từ. Tốt hơn Fasttext ở khía cạnh ngữ nghĩa
  - Fasttext: quan tâm đến cấu trúc từ tách từ thành các âm tiết. Tốt hơn gensim ở khía cạnh cú pháp
- Nhược điểm:
  - Không thể hiện được sự đại diện theo ngữ cảnh cụ thể của từ trong từng lĩnh vực hay văn cảnh cụ thể
    BERT

word2vecVN



Search
giận .*
by
label
13 matches.
tức_giận
giận
giận_dữ
nổi_giận
nóng_giận
giận_dỗi
giận_hờn
cáu_giận
chọc_giận
hờn_giận
hả_giận
căm_giận
oán_giận

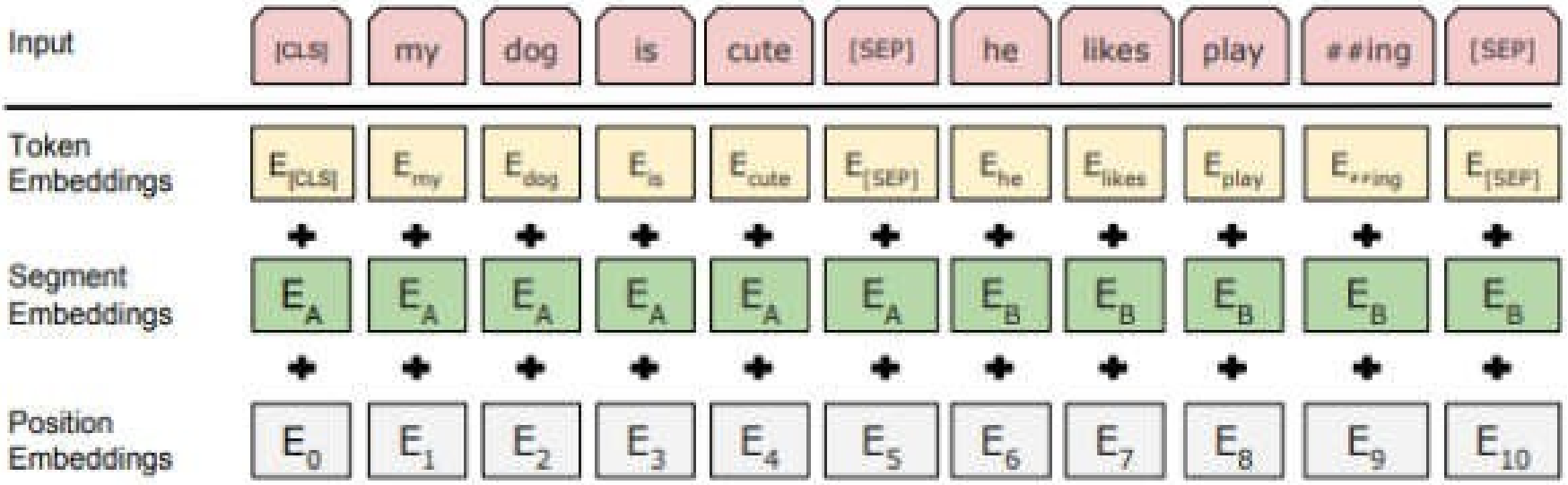
https://github.com/sonvx/word2vecVN

Bidirectional Encoder Representations from Transformers (BERT)

- Bert là mô hình biểu diễn ngôn ngữ của Google, sử dụng pre-training and fine-tuning để tạo ra các mô hình hiện đại cho nhiều tác vụ: Question Answering, sentiment analysis,…..
- BERT huấn luyện thông qua ngữ cảnh 2 chiều của Transformer

BERT

- Input: 1 câu hoặc 1 cặp câu (ví dụ: [Câu hỏi, câu trả lời])

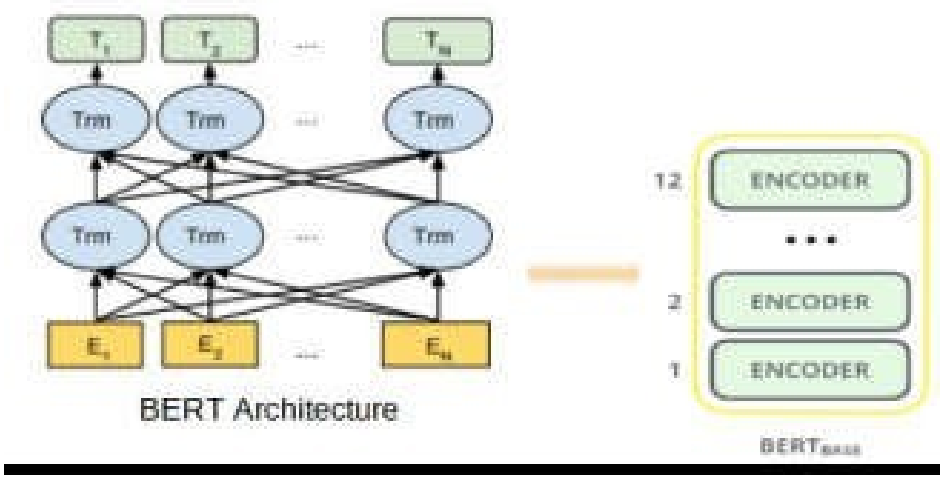


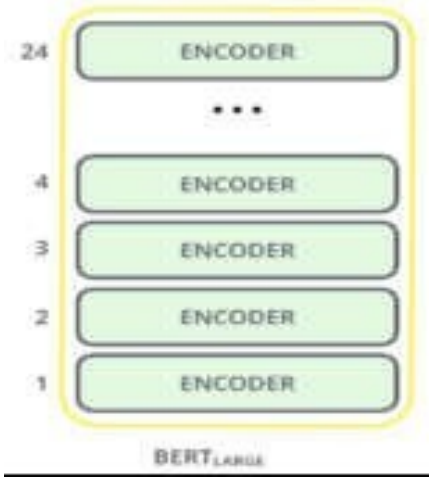
BERT

- **Positional embeddings:** vị trí token trong câu, tối đa 512 tokens.
- **Token embeddings**: các token của xâu đầu vào. Token đầu tiên là [CLS]. Token kết thúc câu là [SEP]. Trong task phân loại, đầu ra của Transformer (hidden state cuối cùng) ứng với token này là giá trị phân loại.
- **Segment embeddings**: phân biệt 2 câu trong trường hợp đầu vào là cặp câu, câu A là các giá trị 0, câu B là các giá trị 1.

Kiến trúc BERT

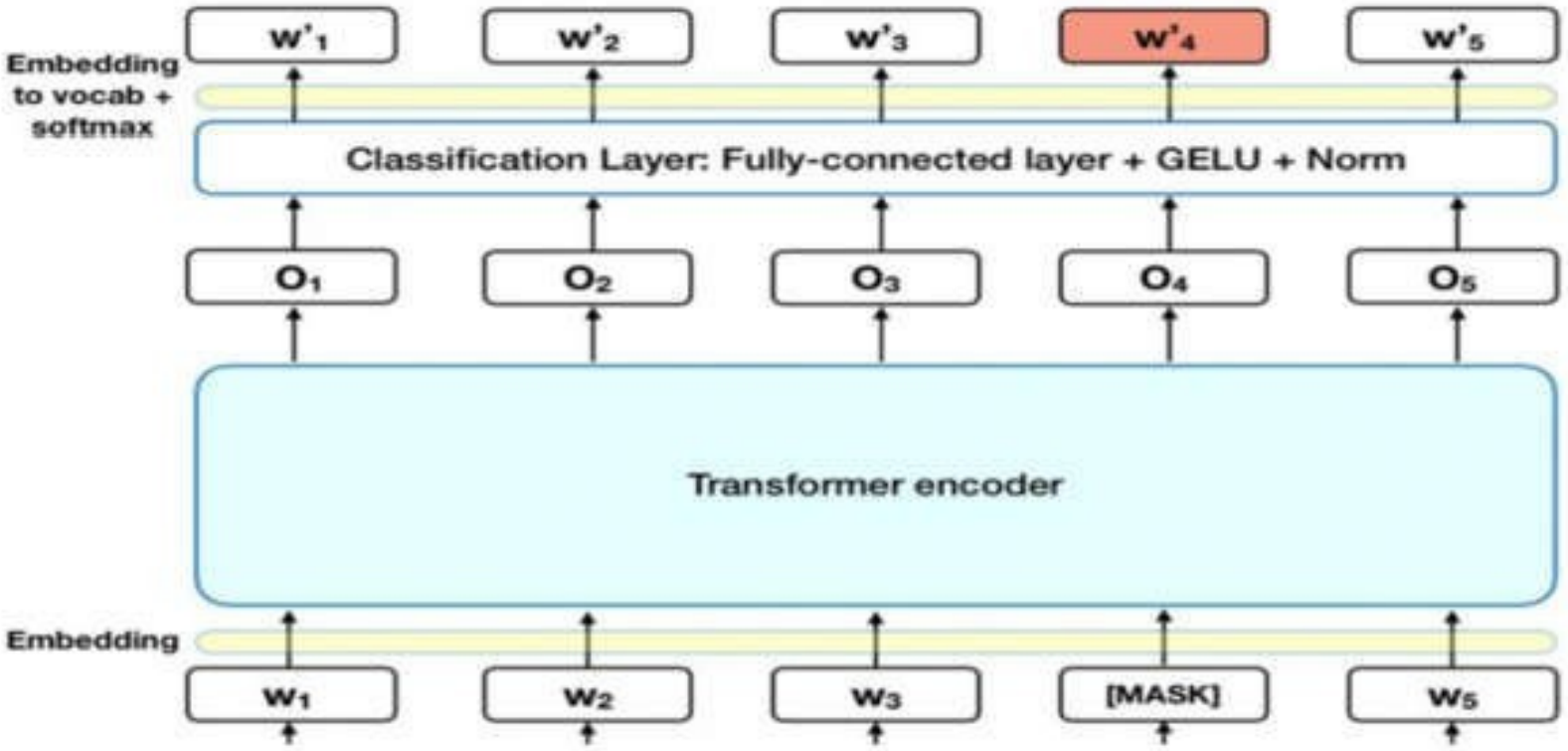
- BERT sử dụng bộ mã hóa Transformer 2 chiều nhiều lớp.
  Lớp self-attention thực thi self-attention theo cả 2 hướng
- Google công bố 2 dạng của mô hình:
  - BERT Base: 12 layers (transformer blocks), 12 attention heads,110M parameters
  - BERT Large: 24 layers (transformer blocks), 16 attention heads, 340M parameters





Huấn luyện BERT

- BERT được pre-training sử dụng 2 tác vụ dự đoán không giám sát
  - Masked Language Modeling (MLM)



Huấn luyện BERT

- **Next Sentence Prediction (NSP)**
  - BERT sử dụng các cặp câu làm dữ liệu train. Ví dụ: sử dụng bộ dữ liệu 100.000 câu để pre-training 1 mô hình ngôn ngữ => có 50.000 mẫu train (cặp câu) làm dữ liệu train
  - Với 50% các cặp, câu thứ 2 sẽ là câu tiếp theo cho câu thứ nhất. Các nhãn này ký hiệu là "IsNext"
  - Với 50% còn lại, câu thứ 2 sẽ là một câu ngẫu nhiên từ bộ dữ liệu. Các nhãn này ký hiệu là "notNext"
- ***Note:*** Khi train mô hình BERT thì MLM và NSP được train cùng nhau để giảm thiểu lỗi

BERT

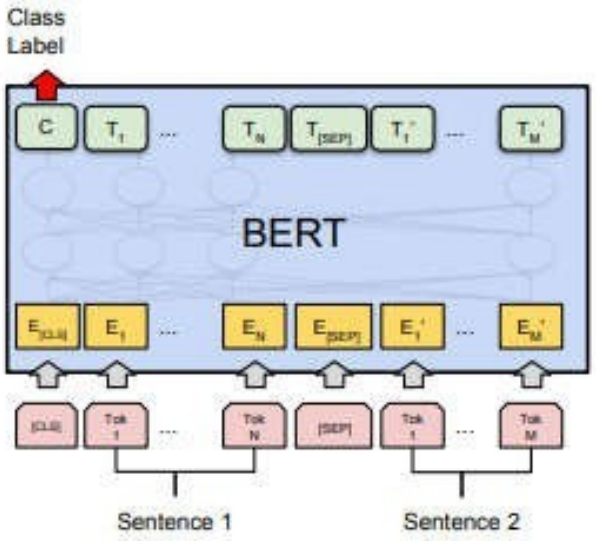
Input: [CLS] người đàn_ông làm [MASK] tại cửa_hàng [SEP] anh_ta rất [MASK] và thân_thiện [SEP]

Label: isNext

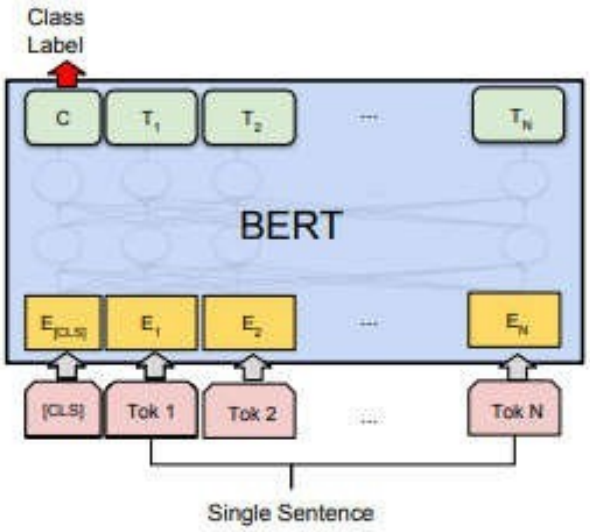
Input: [CLS] người đàn_ông làm [MASK] tại cửa_hàng [SEP] cô_ta đang cầm súng [SEP]

Label: notNext

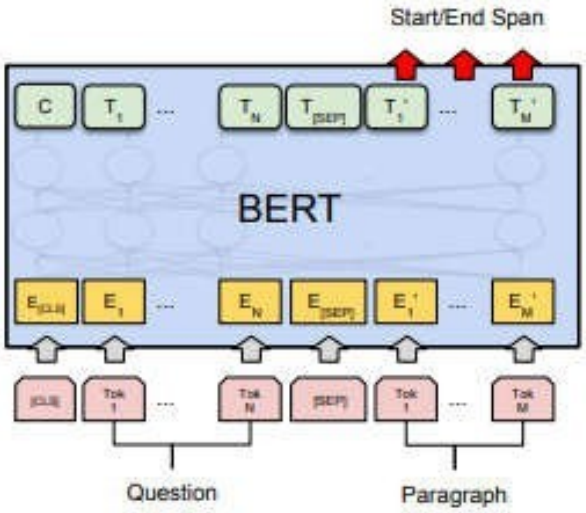
Một số mô hình sử dụng BERT



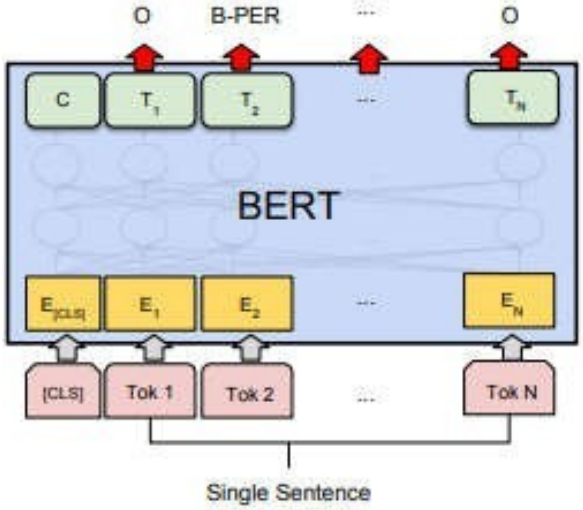
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



(b) Single Sentence Classification Tasks: SST-2, CoLA



(c) Question Answering Tasks: SQuAD v1.1



(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

WSD và IR

- IR (Information Retrieval) : tìm kiếm thông tin
- Motivation
  - Đồng âm = Bank (ngân hàng, sông)
  - Đa nghĩa = Bat ((câu lạc bộ chơi cricket), (cây vợt nhỏ có tay cầm dài để chơi bóng))
  - Đồng nghĩa = doctor, doc, physician, MD, medico
- Những vấn đề trên ảnh hưởng đến IR như thế nào?
  - Đồng âm và đa nghĩa có xu hướng giảm độ chính xác
  - Đồng nghĩa: giảm độ phủ

2 ứng dụng của WSD trong IR

- Tìm kiếm dựa trên câu truy vấn (Voorhees, 1998):
  - Sử dụng WSD để mở rộng câu truy vấn: phân giải nhập nhằng câu query và bổ sung vào các từ có nghĩa rộng hơn.
  - Sử dụng WSD để đánh chỉ số khái niệm: phân giải nhập nhằng tập tài liệu và xây dựng chỉ số cho tập synset thay vì cho tập từ gốc
  - Mô hình không gian vector: tìm độ tương đồng cosin giữa câu truy vấn và mỗi vector tài liệu
- Đánh chỉ số khái niệm
  - Trong các thí nghiệm, vector dựa trên nghĩa thực hiện kém hơn vector dựa trên từ gốc
  - Lý do: lỗi phân giải nhập nhằng
    - trong thu thập văn bản, và
    - các câu query ngắn do thiếu nội dung

2 ứng dụng của WSD trong IR

- Mở rộng query
  - Không khả quan
  - Nhưng, phân giải nhập nhằng và mở rộng truy vấn thủ công đem lại kết quả tốt
- Ví dụ:
  - *furniture*: table, chair, board, refectory(specialisations)
  - “Chỉ có một vài từ vựng liên quan là có ích trong việc mở rộng câu truy vấn, vì đường dẫn lớp cha giữa các từ trong WordNet không phải lúc nào cũng đem lại 1 mở rộng truy vấn 1 cách hữu ích”

Độ chính xác của WSD và IR

- Tập dữ liệu đánh giá WSD: SensEval và SemCor
- Cách khác để tạo ra dữ liệu gán nhãn: Pseudowords
  - Lấy 2 từ (ngẫu nhiên) có cùng từ loại, và thay thế cả 2 bằng 1 từ nhân tạo. Ví dụ, 'door' và 'banana' có thể thay thế trong tập ngữ liệu bằng từ 'donana'.
  - Độ chính xác của WSD: xác định được mỗi trường hợp của donana cụ thể là 'door‘ hay 'banana'. (Yarowsky, 1993)
- (Sanderson, 1997) công bố: thêm nhập nhằng vào các query và kết quả ít có ảnh hưởng đến độ chính xác của việc tìm kiếm so với ảnh hưởng của lỗi phân giải nhập nhằng trong tập kết quả
  - chỉ có lỗi phân giải nhập nhằng mức thấp (< 10%) mới tốt hơn phiên bản IR đơn giản dựa trên từ gốc.

Độ chính xác của WSD và IR

- Tại sao đa nghĩa/đồng âm không phải vấn đề lớn như ta nghĩ:
  - Tác động của sự đồng xuất hiện từ truy vấn: các từ trong câu truy vấn tự nó đã phân giải nhập nhằng
  - Sự phân bố ngữ nghĩa: áp dụng cho các miền ứng dụng cụ thể

Độ chính xác của WSD và IR

- Từ đồng nghĩa có ảnh hưởng lớn hơn:
  - Gonzalo et al. (1998; 1999): sử dụng SemCor (tập ngữ liệu Brown với các thẻ nghĩa của WordNet) cho thấy nếu phân giải nhập nhằng có độ cx = 100%
    - Đánh chỉ số nghĩa (vd synset number) có độ cx IR = 62%
    - Đánh chỉ số nghĩa của từ (vd canine1) có độ cx IR = 53.2%
    - Đánh chỉ số từ gốc có độ cx IR = 48%
  - Gonzalo et al. cho thấy độ cx tối thiểu 90% với WSD cho IR là quá cao. Gần 60% từ giả không hoạt động giông như từ có nhập nhằng thật.

Bài tập

1. Tìm kiếm 1 mã nguồn/thư viện LDA trên github, cài đặt và chạy thử với 1 tập ngữ liệu tiếng Anh:
   - Đặt số topic là 10. In ra màn hình các từ khóa đại diện cho từng topic.
2. Tìm kiếm 1 mã nguồn/thư viện word2vec cho tiếng Việt trên github, cài đặt và chạy thử:
   - Tìm từ đồng nghĩa với một số từ cho trước
   - Đo độ tương đồng giữa 2 câu tiếng Việt