**Question classification using Head words and their Hypernyms**

**1. Tổng quan của bài báo :**

Phân lớp câu hỏi có vai trò quan trọng trong hệ thống giải đáp (QA system) . Để phân loại chính xác cần dựa trên một số các đặc trưng phân loại (features) . Bài báo đưa ra đặc trưng phân loại dựa trên Head Word , và trình bày hai hướng tiếp cận để gia tố thêm đặc trưng ngữ nghĩa của Head word bằng cách sử dụng Wordnet.

Bài báo còn đề cập tới một thuật toán dùng để giải quyết nhập nhằng ngữ nghĩa của [Michael E. Lesk](http://en.wikipedia.org/wiki/Mike_Lesk) -1986 (word sense disambiguation – WSD) . Thuật toán WSD được tác giả sử dụng một cách thích hợp với ngữ cảnh của bài báo này và độ sâu của đặc trưng hypernym (trong Wordnet) được tối ưu hóa .

Với thêm một số đặc trưng chuẩn gia tăng thêm như unigram , thì hai mô hình linear SVM và Maximum Entropy (ME) đạt độ chính xác đến 89.2% và 89.0% tương ứng với tập dữ liệu chuẩn benchmark .

**2. Bộ phân loại (Classifier):**

Bộ phân loại bài báo sử dụng để đánh giá là Support Vector Machine (SVM) và Maxium Entropy (ME) . Hai bộ phân loại này trong việc phân loại câu hỏi có thể hiện gần giống như nhau (dựa trên đánh giá kết quả của bài báo ).

**3. Support Vector Machine**

**4.Các đặc trưng phân loại ( Features)**

Trong bài báo này, tác giả đề ra 5 đặc trưng phân loại : wh-word , word head , Wordnet semantic features for head word , word grams ,word shape . Mỗi đặc trưng này sẽ được sử dụng riêng bởi bộ phân loại để đánh giá được mức độ đóng góp của từng đặc trưng.

**4.1. Question wh-word :**

Đặc trưng này dựa trên những từ thuộc loại wh-word trong câu hỏi . Ví dụ : *What is the population of China* thì từ wh-word là *What* . Một tập 8 từ wh-word bao gồm : what, which, when, where, who, how, why và từ khác . Từ khác là từ mà không thuộc vào những loại từ nào trước nó . Ví dụ : *Name a food high in zinc* .

**4.2. Head word**

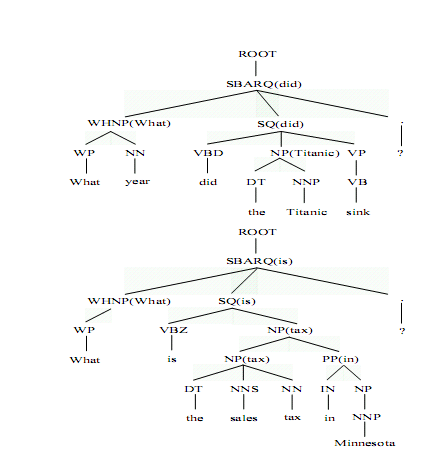
“*head word, which is one single word specifying the object that the question seeks”* (line 9 page 929 “Question classification using Head words and their Hypernyms”)

Như khái niệm trên , head word là một từ đặc tả cho đối tượng mà câu hỏi đang tìm kiếm . Ví dụ : *what is a group of turkeys called* có từ head word chính xác là *turkeys* .

Với một ví dụ khác : *George Bush purchased a small interest in which baseball team* , theo cách tiếp cận dựa trên Head chunk của Li và Roth(2002,2006) và informer span của Krisman (2005) thì kết quả sẽ cho ra là *baseball team* thay vì từ *team* mà đặc trưng Head word trả về . Từ phụ *baseball* sẽ dẫn đến nhầm lẫn trong việc phân loại theo ENTY:sport hơn là phân loại theo HUM:group.

Để thực hiện việc lấy head word , một bộ phân tích cú pháp cần được sử dụng . Một số bộ phân tích cú pháp chuẩn xác mà bài báo đưa ra gồm : Chaniak parser, Stanford parser , và Berkeley parser . Bài báo này chọn Berkeley parser để sử dụng.

Sau khi phân tích cú pháp ,ta đươc một cây cú pháp . Luật của Collins (1999) (tham khảo phụ lục 1) được áp dụng trên cây cú pháp này để rút ra những head word dựa trên cú pháp .

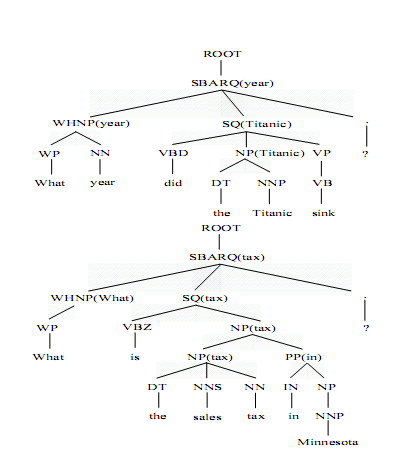


Hình 1 : rút head word dựa trên cây cú pháp

(Figure 1 of “Question classification using Head words and their Hypernyms”,2008 )

Theo hình trên tương ứng với hai câu hỏi : *What year did the Titanic sink* và *What is the sales tax in Minnesota respectively* . Theo luật của Collins thì trong ngữ WHNP của câu đầu thì phần tử con WP được lấy làm head word ( vì thế từ *what*  nằm trong ngoặc đơn ) và từ *did* được rút ra là head word cho toàn bộ câu . Tương ứng cho câu thứ hai là từ *is*.

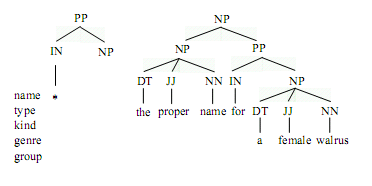
Với nhu cầu của tác giả , những luật của Collins (SBARQ,SQ,VP,SINV)được thay đổi cho phù hợp : từ head sẽ tham chiếu tới danh từ (hoặc cụm danh từ) thay vì động từ (hoặc cụm động từ ). Với thay đổi như thế thì kết quả cho ra như hình bên dưới (hình 2)



Hình 2 – Rút head word dựa trên cây cú pháp sau khi định nghĩa lại các luật

(Figure 2 of “Question classification using Head words and their Hypernyms”,2008 )

Tuy nhiên , nếu head word nằm thuộc trong số các từ : name, kind, type ... thì cẩn phải được điểu chỉnh lại cho đúng head word thật sự . Một cây cú pháp mẫu được đưa ra để điều chỉnh lại (Hình 3 )



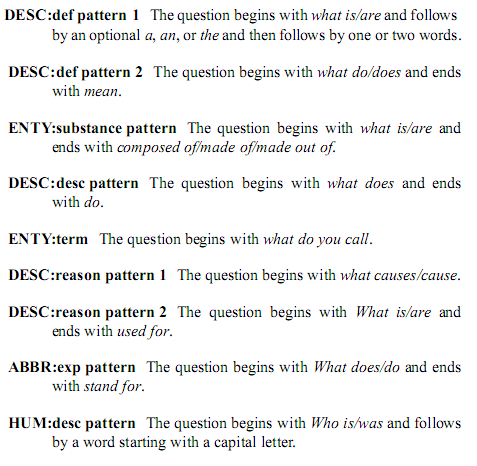
Hình 3 – cây cú pháp điều chỉnh head word

(Figure 3 of “Question classification using Head words and their Hypernyms”,2008 )

Nếu mẫu cây trên khớp với cây cú pháp của câu hỏi được đưa ra , thì head word sẽ được chỉ định lại nằm ở nhánh NP theo mẫu . Ví dụ : *What is the proper name for a female walrus .* Theo cách trên thì head word sẽ là *name* .Nhưng sau khi được chỉ định lại thì head word chính xác sẽ là *walrus* ( tương ứng với lớp ENTY:animal)

Ngoài ra , tác giả đề cử ra một số pattern dựa trên regular expression để hỗ trợ thêm trong quá trình nhận dạng head word . ví dụ : *What is an atom* và *What are intvertebrates* , có những head word là *atom* và *invertebrates* . Những từ này không thể phân loại hai câu trên vào mục DESC:def .

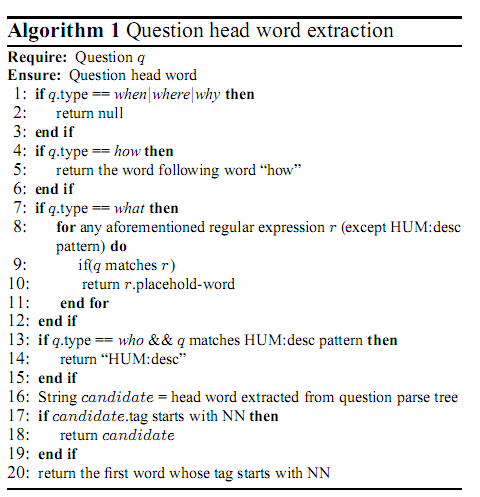
Để giải quyết cho ví dụ trên , Tác giả tạo ra một số chuỗi regular expression (được coi như là một đặc trưng nhị phân ) . Tác giả đưa ra 9 dạng Pattern như hình dưới :



Hình 4 – 9 biểu thức thông dụng dùng cho phân loại câu hỏi

(page 930-931 “Question classification using Head words and their Hypernyms”,2008 )

Cuối cùng , tác giả đưa ra thuật toán ( hình 5 ) cho tất cả những gì trình bày bên trên để rút ra head word cho một câu hỏi .



Hình 6 – thuật giải rút ra head word

(page 931 “Question classification using Head words and their Hypernyms”,2008 )

**4.3. Wordnet Semantic Feature**

Wordnet là một công cụ hữu dụng trong phân tích ngữ ngữ nghĩa và sử dụng rộng rãi trong phân loại câu hỏi ( Krishnan 2005 , Schlaefer 2007) . Hypernym trong wordnet là một đặc điểm mà người dùng hay sử dụng . Trong bài báo này , tác giả đề ra hai hướng để gia tố thêm cho đặc điểm phân loại dựa trên hypernym (Wordnet).

**4.3.1 Sử dụng trực tiếp Hypernym :**

Trong wordnet , các ngữ nghĩa được tổ chức theo một cấu trúc có thứ bậc dựa trên mối quan hệ hypernym .

Ví dụ : cấu trúc thứ bậc cho danh từ *dog*

*Dog -> domestic animal -> animal*

*Dog -> unpleasant woman ->unpleasant person*

Ngoài ra , đối với động từ của *dog*

*Dog -> pursue ->travel*

Ở đây ta đề cập tới head word rút ra ở đặc trưng bên trên , điều này có thể gây nhiễu khi áp dụng hypernym cho head word . Tác giả đặt ra ba câu hỏi cho vấn đề này như sau :

*1. Which part of speech senses should be augmanted ?*

*2. Which sense of the given word is needed to be augment?*

*3. how many depth are required to tradeoff the generality and the specificity ?*

**Câu hỏi đầu tiên** đề cập tới việc phân loại từ (gán nhãn từ ) . Việc này được giải quyết khá dễ dàng : head word sẽ được đưa vào wordnet để gán nhãn từ loại , do wordnet có hỗ trợ chức năng này ( dựa trên chuẩn gán nhãn từ loại Penn Treebank).

**Câu hỏi thứ hai** liên quan tới vấn đề giải quyết nhập nhằng ngữ nghĩa (WSD) theo thuật toán của Lesk. Nó dựa trên giả định rằng các từ trong một ngữ ảnh sẽ có khuynh hướng chia sẽ cùng một chủ đề chung . Cách thực hiện được nêu ra như sau :

B1 : Chọn ra một cặp từ nhập nhằng ngữ nghĩa có cùng ngữ cảnh

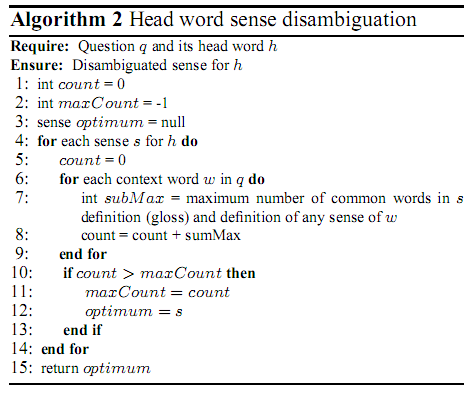
B2 : Kiểm tra định nghĩa của chúng trong một từ điển

B3 : chọn những nghĩa nào mà làm tăng tối đa số lượng từ phổ biến trong định nghĩa của hai từ đã chọn .

(page 932 “Question classification using Head words and their Hypernyms”,2008 )

Công việc chúng ta là giải quyết nhập nhằng ngữ nghĩa của head word . Những từ ngữ cảnh sẽ là những từ có trong câu hỏi ( ngoại trừ head word ) . Từ điển cụ thể là một lời giải thích về ý nghĩa của một từ đưa ra .

Tác giả đã dựa trên thuật toán trên của Lesk để đề ra giải thuật giải quyết nhập nhằng ngữ nghĩa cho Head Word.

**

Hình 7 – thuật toán giải quyết nhập nhằng ngữa nghĩa cho Head word

(page 931 “Question classification using Head words and their Hypernyms”,2008 )

Với mỗi nghĩa của head word , thuật toán tính ra tối đa số lượng từ thông dụng giữa lời giải thích của nghĩa hiện tại và các lời giải thích cho các từ ngữ cảnh . Trong số tất cả các nghĩa của head word , nghĩa nào có tối đa số lượng từ thông dụng sẽ được chọn .

Với câu hỏi cuối cùng , tác giả dựa trên các thử nghiệm và đánh giá đã đưa ra độ sâu là 6 . Với độ sâu này, nó sẽ cho ra kết quả tốt nhất và nó sẽ không làm nhiễu không gian đặc trưng (feature space).

**4.3.2 Gián tiếp sử dụng Hypernym**

Ở hướng tiếp cận này , tác giả sử dụng một gói của Wordnet gọi là WordNet Similarity . Ta đưa vào một cặp từ , gói này sẽ mô hình hóa chiều dài đường đi của một từ đến một từ qua mạng lưới Wordnet, sau đó tính ra độ tương tự ngữ nghĩa .

Vỉ dụ : độ tương tự giữa từ *car* và *automobile* là 1.0 , giữa từ *film* và *audience* là 0.38

Với cách này , ta tính được độ tương tự giữa head word và mỗi từ miêu tả (description word ) trong mục phân lớp . Từ miêu tả cho một mục phân loại thông thường có từ một tới ba từ .( tham khảo phục lục 2 )

Ví dụ : trong mục *ENTY:dismed* có từ miêu tả là *diseases* và *medicine* .

**4.4 N-grams**

“N-gram is a sub-sequence of N words from a given question” (Zhiheng Huang,Marcus Thint , ZhengChang Qin ,2008). Trong cách tiếp cận dựa trên đặc trưng N-gram , ta có 3 dạng :

+Unigram(bag of words)

+Bigram : lấy một cặp từ làm đặc trưng

+Trigram : lấy một bộ 3 tử liên tiếp.

Trong bài báo này , tác giả sử dụng cả 3 dạng trên để đánh giá . Mục đích của việc áp dụng đặc trưng này, theo tác giả , là để phục vụ cho việc giải quyết nhập nhằng ngữ nghĩa .

**4.5 Word Shape**

Đặc trưng này dựa trên hình dạng của từ . Ví dụ : *“Who is Duke Ellington”* có dạng kết hợp bắt đầu bằng một ký tự viết hoa theo sau là những ký tự thường , từ *“Duke”* sẽ được nhận diện qua bộ nhận biết thực thể đặt tên .

Tác giả sử dụng 5 đặc trưng sau :

+all upper case : viết hoa tất cả

+all lower case : viết thường tất cả

+mixed case : kết hợp viết hoa lẫn thường

+all digits : tất cả đều là số

+ other : những trường hợp còn lại

**5.Đánh giá**

Trong việc đánh giá , bao gồm hai giai đoạn : Đánh giá sự đóng góp riêng lẻ của từng đặc trưng, đánh giá sau khi gia tăng đặc trưng . Tập dữ liệu dùng trong quá trình huấn luyện gồm 5500 câu hỏi của UIUC .

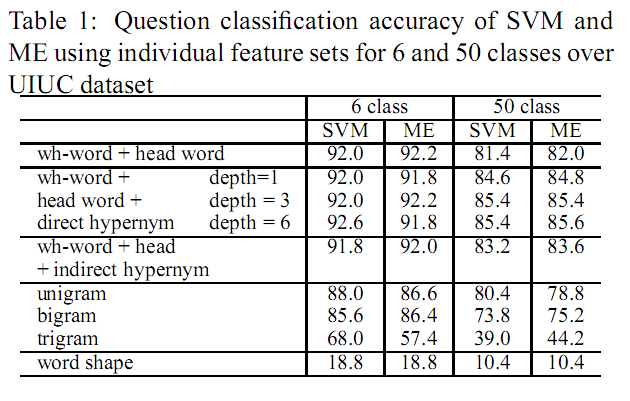
**5.1 Đánh giá từng đặc trưng**

Ở phần đánh giá này , tác giả đánh giá trên các đặc trưng sau :

1. wh-word + head-word
2. wh-word + head-word + direct hypernym
3. wh-word + head-word + indirect hypernym
4. unigram
5. bigram
6. trigram
7. word shape

Các đặc trưng 1,2,3 được đặt ra dựa trên thực tế là wh-word và head word được xem như là một đơn vị(đều dựa trên một đơn vị từ trong câu) và hypernym phục thuộc vào head word .

Bảng kết quả đánh giá :



Hình 8 – Kết quả đánh giá từng đặc trưng

( Table 1 “Question classification using Head words and their Hypernyms”,2008 )

Theo bảng trên , ta có thể thấy rằng việc áp dụng đặc trưng wh-word và head word cho độ chính xác khá tốt ( 50 Class ; SVM 81.4% , ME 82%) . Nhưng nếu kết hợp thêm đặc trưng hypernym thì độ chính xác tăng lên khoảng 4 %. Ở đây , với đặc trưng thứ 2 , có 3 mức độ đánh giá dựa vào độ sâu của hypernym ( Depth =1 , Depth =3 , Depth =6) . Với độ sâu là 6 thì cho ra kết quả tốt nhất ( 50 Class ; SVM 85.4% , ME 85.6%)

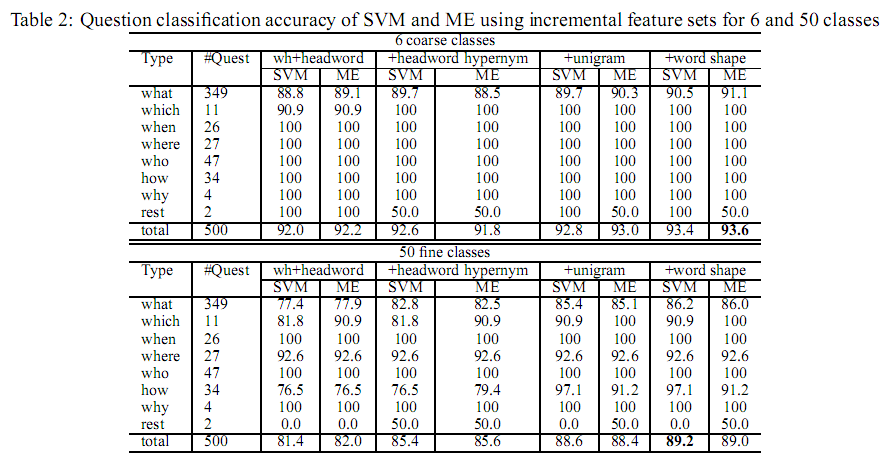
Với việc gián tiếp xài hypernym , kết quả cho ra không tốt bằng việc sử dụng trực tiếp . Có hai lý do mà tác giả nêu ra cho kết quả này là :

1. Từ miêu tả (description word ) chỉ khoảng một tới ba từ , không đủ đại diện cho một mục phân loại .
2. Việc sử dụng gián tiếp không hiệu quả bằng sử dụng trực tiếp .

Trong ba đặc trưng unigram,bigram và trigram thì unigram có được kết quả tốt nhất . Còn với đặc trưng word shape thì độ chính xác không cao . Nó không cung cấp đủ thông tin để phân lớp .

**5.2 Đánh giá sau khi gia tăng đặc trưng**

Dựa vào kết quả đánh giá ở trên , các đặc trưng mà tác giả dùng để huấn luyện cho SVM và ME là : wh-word,head-word,direct hypernym of head word (depth =6),unigram và word shape . Bảng dưới đây cho sẽ cho ta thấy kết quả :



Hình 9 – Kết quả khi kết hợp các đặc trưng

( Table 2“Question classification using Head words and their Hypernyms”,2008 )

Nhìn vào hình trên cho thấy kết quả phân lớp tốt nhất đối với 50 phân lớp nhỏ là 89.2%(SVM) và đối với 6 phân lớp lớn là 93.6%(ME) .Tuy nhiên vẫn còn 10% xác xuất lỗi trong việc phân lớp 50 phân lớp con . Tác giả nếu ra 3 lý do :

1. Sự nhập nhằng tồn tại trong việc phân loại một câu hỏi.

Vd : *What is mad cow disease* (DESC:def or ENTY:dismed)

1. Ghi nhãn không phù hợp trong dữ liệu đào tạo và kiểm tra.

Vd : *What is country is Chicago in* (LOC:other)

*What country is Phoenix ,AZ in* (LOC:city)

1. Bộ phân tích cú pháp có thể cho ra kết quả sai cho cây cú pháp.

Vd : *What is the speed hummingbirds fly*

(head word là *hummingbirds* (ENTY:animal) thay vì từ chính xác là *speed* (NUM:speed)

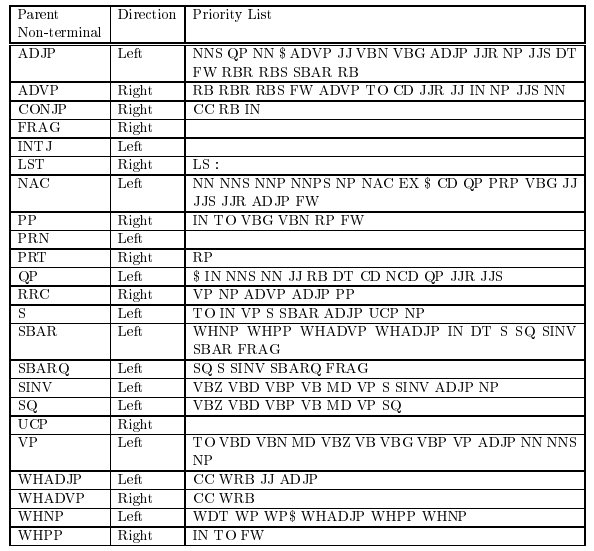
**6.Phụ lục**

**6.1 Phụ lục 1 :**

**Collins rules** là luật được đưa ra để tìm những từ chính trong các thành phần của một cây cú pháp . Các luật bên dưới được tham khảo từ phục lục A trong luận văn của Collins “Head-Driven Statistical Models for natural language parsing” 1999)

*Cách sử dụng bảng* : các luật được biểu diễn dưới dạng <X -> Y1...Yn> .

* Trong đó X : cụ thể là một VP (một Parent Non-Terminal trong hình dưới )
* Thuật toán đầu tiên sẽ tìm từ phía bên trái của dãy <Y1 ... Yn> , tìm Yi nào là TO , nếu không tìm thấy sẽ tìm từ tiếp theo là VBP và cứ tiếp tục như thế .
* Nếu chạy hết dãy mà không tìm thấy từ nào , thì phần tử con nằm ngoài cùng bên trái sẽ được chọn , tức Y1 .

****

Hình 10 – Bảng danh sách các luật

(Table A.1 Collins “Head-Driven Statistical Models for natural language parsing” 1999)

*Luật riêng cho NPs*

* Nếu là từ cuối cùng được gán nhãn (tagged POS), trả về từ cuối
* Nếu không , tìm từ *phải* sang *trái* phần tử con là NN , NNP , NNPS,NNS ,NX,POS và JJR .
* Nếu không , tìm từ *trái* sang *phải* phần tử đầu tiên nào là NP.
* Nếu không , tìm từ *phải* sang *trái* phần tử đầu tiên nào là $ , ADJP hoặc PRN
* Nếu không , tìm từ *phải* sang *trái* phần tử đầu tiên nào là CD.
* Nếu không , tìm từ *phải* sang *trái* phần tử đầu tiên nào là JJ, JJS ,RB hoặc QP .
* Nếu không , trả về từ cuối cùng .

**6.2 Phụ lục 2 : Description word**

Bảng định nghĩa cho 6 phân lớp lớn và 50 phân lớp nhỏ . (<http://l2r.cs.uiuc.edu/~cogcomp/Data/QA/QC/definition.html> )

|  |  |
| --- | --- |
| **Class** | **Definition** |
| ***ABBREVIATION***  **abb**  **exp** | Abbreviation  Abbreviation  Expression abbreviated |
| ***ENTITY***  **animal**  **body**  **color**  **creative**  **currency**  **dis.med.**  **event**  **food**  **instrument**  **lang**  **letter**  **other**  **plant**  **product**  **religion**  **sport**  **substance**  **symbol**  **technique**  **term**  **vehicle**  **word** | Entities  Animals  Organs of body  Colors  Inventions, books and other creative pieces  Currency names  Diseases and medicine  Events  Food  Musical instrument  Languages  Letters like a-z  Other entities  Plants  Products  Religions  Sports  Elements and substances  Symbols and signs  Techniques and methods  Equivalent terms  Vehicles  Words with a special property |
| ***DESCRIPTION***  **definition**  **description**  **manner**  **reason** | Description and abstract concepts  Definition of sth.  Description of sth.  Manner of an action  Reasons |
| ***HUMAN***  **group**  **ind**  **title**  **description** | Human beings  A group or organization of persons  An individual  Title of a person  Description of a person |
| ***LOCATION***  **city**  **country**  **mountain**  **other**  **state** | Locations  Cities  Countries  Mountains  Other locations  States |
| ***NUMERIC***  **code**  **count**  **date**  **distance**  **money**  **order**  **other**  **period**  **percent**  **speed**  **temp**  **size**  **weight** | Numeric values  Postcodes or other codes  Number of sth.  Dates  Linear measures  Prices  Ranks  Other numbers  The lasting time of sth.  Fractions  Speed  Temperature  Size, area and volume  Weight |