Hiện nay, trong tiếng Anh, câu hỏi được [2,3] chia làm 3 lọai cơ bản là: Yes/No Question, Question Word Question và Choice Question. Đây là các lọai câu hỏi thông dụng mà ta thường thấy. Ngòai ra, ta còn có thể thấy một số lọai câu hỏi khác như tag question, open question, closed question, direct/ indirect question [1,4].

Với tập dữ liệu TREC, các câu hỏi lại được phân làm 3 lọai : FACTOID, LIST và OTHER. Câu hỏi thuộc lọai FACTOID , theo như tham khảo thì nó là một loại câu hỏi dành cho những vấn đề chưa được xác minh về thông tin và không chắc là có thông tin để trả lời. Câu hỏi LIST là câu hỏi thuộc dạng liệt kê một danh sách theo yêu cầu của câu hỏi và còn lại là các câu hỏi OTHER.

Tuy nhiên, đa số các nghiên cứu về hệ thống hỏi đáp thì không phân loại như vậy. Đối với một số nghiên cứu của các bài báo [5-8]về việc phân lọai câu hỏi thì câu hỏi thuộc phân lọai theo 6 mục thô và 50 mục mịn. Các mục này được dưa ra bởi Li và Roth (2002). Đây là hình thức phân lọai mà được nhiều bài báo lựa chọn.

|  |  |
| --- | --- |
| Mục thô | Mục mịn |
| ABBR | abbreviation,expansion |
| DESC | definition, description, manner, reason |
| ENTY | animal, body, color, creation, currency, disease, event, food, instrument, language, letter, other, plant, product, religion, sport, subtance, symbol, technique, term, vehicle, word |
| HUM | description, group, individual, title |
| LOC | city, country, mountain, other, state |
| NUM | Code, count, date, distance, money, order, other, percent, period, speed, temperature, size, weight |

Bảng 1 – Phân loại 6 mục thô và 50 mục mịn theo Li và Roth

Vậy, tùy theo mục đích sư dụng mà câu hỏi còn đựơc phân lọai theo nhiều mục. Đối với 3 lọai câu hỏi cơ bản trong tiếng Anh thì hệ thống của nhóm hiện tại không giải quyết các câu hỏi yes/no question và choice question. Đó là các câu hỏi thuộc dạng xác minh vấn đề (có hoặc không) và câu hỏi lựa chọn. Hệ thống chỉ giải quyết được một số các câu hỏi thuộc loại Question Word Question.

Các câu hỏi đặt ra trong đề tài này sẽ được phân loại theo 6 mục thô (ABBR, DESC, ENTY, HUM, LOC, NUM). Cách phân loại này gần với đề tài hơn vì nó cùng liên quan đến hệ thống hỏi đáp. Bởi vì như đã trình bày, hệ thống chủ yếu xử lý các câu hỏi dạng Question Word Question, tức là có các từ để hỏi như: what, who, when, where… Vì thế, nhóm sẽ phân lọai các câu hỏi trong hệ thống theo 6 mục thô của Li và Roth (2002). Việc phân lọai theo 50 mục mịn là không cần thiết, vì số lượng câu hỏi giải quyết được trong đề tài chỉ thuộc về một phạm vi xung quanh các bài báo công nghệ thông tin. Hơn nữa, số lượng câu hỏi không nhiều để phân vào 50 mục mịn. Việc phân lọai sẽ được phân lọai bằng tay, theo ý kiến của các thành viên trong nhóm.

Nguồn câu hỏi được tham khảo từ :

- Dựa trên nguồn dữ liệu của DBLP ( các thuộc tính trong các table)

- Nhu cầu người dùng : các câu hỏi liên quan đến vấn đề mà người dùng cần biết.

- Một số các mẫu câu trong tập TREC 2003 và 2007

|  |  |  |
| --- | --- | --- |
| **Phân loại** | **Câu hỏi** | **Kết quả** |
| ENTY | 1. Which books were written by Rafiul Ahad and Amelia Carlson ? | *1.Object SQL - A Language for the Design and Implementation of Object Databases.(1995)* |
| ENTY | 1. Which books were written by Rafiul Ahad and Amelia Carlson in 1995 ? | *1.Object SQL - A Language for the Design and Implementation of Object Databases.(1995)* |
| ENTY | 1. Which books were written by Rafiul Ahad from 1999 to 2010 ? | *1. A Scalable and Highly Available Networked Database Architecture.(1999)* |
| ENTY | 1. Which books were published by O'Reilly in 1999 ? | *1. Java Power Reference (1999)*  *2. Java Foundation Classes in a Nutshell (1999)*  *3. Java 2D Graphics (1999)*  *4. Windows 98 in a Nutshell (1999)*  *5. CJKV Information Processing: Chinese, Japanese, Korean & Vietnamese Computing (1999)*  *6. Programming the Be Operating System: Writing Programs for the Be Operating System (1999)* |
| NUM | 1. How many papers were written by Rafiul Ahad ? | *11 papers were written by Rafiul Ahad.*  *1. A Scalable and Highly Available Networked Database Architecture.(1999)*  *2. Object SQL - A Language for the Design and Implementation of Object Databases.(1995)*  *3. RQL: A Recursive Query Language.(1993)*  *4. Supporting Access Control in an Object-Oriented Database Language.(1992)*  *5. Using a Relational Database System to Support Explanation in a Knowledge-Based System.(1992)*  *6. …* |
| NUM | 1. How many papers were written by Rafiul Ahad in 1999 ? | *1 papers were written by Rafiul Ahad*  *1.* *A Scalable and Highly Available Networked Database Architecture.(1999)* |
| HUM | 1. Who write books in 1999 ? | *-James Gosling*  *-Ken Arnold*  *-Berthier A. Ribeiro-Neto*  *-Ricardo A. Baeza-Yates*  *…* |
| HUM | 1. Who write books from 1999 to 2010 ? | *-Leslie Lamport*  *-Arthur J. Bernstein*  *-Michael Kifer*  *…* |
| HUM | 1. Who published books from 1999 to 2000 ? | *-CSREA Press*  *-Metropolis*  *-World Scientific*  *-CSLI Publications*  *-Vieweg*  *-ACM*  *…* |
| HUM | 1. Who published books 1999 ? | *-AAAI*  *-AAAI Press*  *-ACM*  *-ACM Press*  *-ACM Press / Addison-Wesley*  *-Addison-Wesley*  *-APress*  *….* |
| ENTY | 1. What are titles of books written by Marcus Thint ? | *1.* *Ranking Answers by Hierarchical Topic Models..(2009)*  *2. Investigation of Question Classifier in Question Answering.(2009)*  *3. Semantic approach to text entailment for question answering - new domain for uncertainty modeling.(2008)*  *4.* *Question Classification using Head Words and their Hypernyms.(2008)*  *…* |
| ENTY | 1. What books did Jennifer Widom write ?   or What books has Jennifer Widom written ? | *1.Synthesizing view definitions from data.(2010)*  *2.* *LIVE: A Lineage-Supported Versioned DBMS.(2010)*  *3.* *Confidence-Aware Join Algorithms.(2009)*  *4.* *Schema Design for Uncertain Databases.(2009)*  *5.* *Swoosh: a generic approach to entity resolution.(2009)*  *…* |
| HUM | 1. Who is the author of “Working Models for Uncertain Data” | *-Anish Das Sarma*  *-Omar Benjelloun*  *-Alon Y. Halevy*  *-Jennifer Widom* |
| ENTY | 1. What book did Philip K. Chan write in 1999? | *1.AdaCost: Misclassification Cost-Sensitive Boosting.(1999)*  *2.* *Constructing Web User Profiles: A non-invasive Learning Approach.(1999)*  *3.Guest Editors' Introduction.(1999)* |
| ENTY | 1. What book did Philip K. Chan write from 1999 to 2000? | *1.AdaCost: Misclassification Cost-Sensitive Boosting.(1999)*  *2.* *Constructing Web User Profiles: A non-invasive Learning Approach.(1999)*  *3.Guest Editors' Introduction.(1999)* |
| ENTY | 1. What are the titles of the books published by O’reilly in 1999 ? | *1. Java Power Reference (1999)*  *2. Java Foundation Classes in a Nutshell (1999)*  *3. Java 2D Graphics (1999)*  *4. Windows 98 in a Nutshell (1999)*  *5. CJKV Information Processing: Chinese, Japanese, Korean & Vietnamese Computing (1999)*  *6. Programming the Be Operating System: Writing Programs for the Be Operating System (1999)* |
| HUM | 1. What composer wrote “ *Java 2D Graphics*” | *-* *Jonathan Knudsen* |
| ENTY | 1. What books has isbn 1-56592-484-3 | *1.Java 2D Graphics (1999)* |
| ENTY | 1. What books has doi 10.1145/360271.360274 | *1.* *Java 2 Distributed Object Middleware Performance Analysis and Optimization.(2000)* |
| HUM | 1. What composer wrote books from 1999 in ACM? |  |
| HUM | 1. Who is the author of the paper “ *Question Classification using Head Words and their Hypernyms.”?* |  |
| HUM | 1. Who wrote “ *Question Classification using Head Words and their Hypernyms.”?* |  |
| ENTY | 1. What books were written by “Philip K. Chan” from ACM? |  |
| NUM | 1. How many publisher did “Philip K. Chan” work with? |  |
| ENTY | 1. What publications have resulted from TREC? |  |

[1] <http://en.wikipedia.org/wiki/Question#Types>

[2] <http://www.englishclub.com/grammar/verbs-questions_types.htm>

[3] <http://oxforddictionaries.com/view/entry/m_en_gb0680290#m_en_gb0680290>

[4] <http://www.learnenglish.de/grammar/questiontext.htm>

[5] Question classification using support vector machines, Dell Zhang and Wee Sun Lee, 2003.

[6] Question Classification by Ensemble Learning, LiXin, Huang Xuan Jing,Wu Lid-de, 2006

[7] Effectiveness of Combined features for machine learning based question classification ,Marcin Skowron and Kenji Araki , 2005

[8] Question classification using Head word and their Hypernyms, Zhiheng Huang, Marcus Thint and Zeng Chang Qin ,2008

[9] Learning Question Classifier, Li and Roth, 2002