# Exploiting Synonyms to Improve Question and Answering Systems

Anietie Andy
Department of Systems and Computer Science,
Howard University

Mugizi Robert Rwebangira
Department of Systems and Computer Science,
Howard University

Mohamed Chouikha
Department of Electrical and Computer Engineering
Howard University

## ABSTRACT

Community Question and Answering (CQA) systems are a popular way for Internet users to get answers to complex and common everyday questions. One of the challenges with CQA is that some of the asked questions are not answered [8]. This paper addresses this challenge by using a synonym-based approach that expands each unanswered question into several related questions. This paper argues that the number of unanswered questions can be reduced by searching the data set for the most similar resolved question(s) (questions that have been satisfactorily answered) to either the unanswered question and / or any of its expanded questions. If this search returns more than one resolved question, we rank the returned questions and choose the highest ranking resolved question as the most similar to the unanswered question.

## General Terms

Information Extraction

## Keywords

Question and Answering Jsystems, resolved question, Yahoo Answers! EXPAND

## 1. INTRODUCTION

Community Question and Answering (CQA) system are systems in which users ask questions in various categories such as Mathematics, Parenting, and Pets to name a few and these questions are answered by other users of the system. Yahoo! Answers, Quora, and Stack Overflow are example of CQA'S [7]. In Yahoo! Answers there are two parts to an asked question:

I) The Subject - is a brief description of the question.

II) The Content - is a more detailed description of the question.

The user who asked the question chooses the answer that best satisfies her question.

A CQA is a good medium to get complex and common questions answered by people who are presumably more knowledgeable about these questions. However, these kinds of systems have some challenges, one of which is that a significant number of asked questions are left unanswered [8]. Various methods to reduce the number of unanswered questions in a CQA have been proposed. One of the proposed methods searches the resolved questions to find the most similar question to an unanswered question. Then the answer to the most similar question is used to satisfactorily answer the unanswered question.

<Subject> How do you fold towels in the shape of animals?</Subject>

<content> usually these towels are found on cruise ships</content>

<best answer> There are examples of a swan and peacock online. There is a book with all kinds of towel animal folding instructions</best answer>

**Figure 1: The subject and content of a question and the chosen best answer for this question.**

A significant number of English words have synonyms. Given a question and the synonyms of the words in this question; the question can be phrased in multiple ways without loosing its meaning by taking the word synonyms into consideration. For example, given the following unanswered question " How do I write a good resume?" this question can be expanded as follows:

I) I need help writing my curriculum vitae?

II) How do I write a good CV?

III) How should I format my resume?

The proposed algorithm contributes the following:

I) Expand Questions: This expands an unanswered question into multiple questions by exploiting word synonyms.

II) Search and rank similar resolved questions to the unanswered question and its corresponding expanded questions.

III) The proposed algorithm uses only the Subject of the question and its expansions. It returns more relevant results than using cosine similarity on the Subject and content parts of the question.

## 2. RELATED WORK

Yahoo! Answers is a website was users post questions and answers, all of which are public to registered users. In Yahoo! Answers, a user asks a question in a predefined category (e.g. Mathematics). This question remains "open" for approximately a week with an option for extension or for less than a week if the asker chooses a best answer within this period. If the asker does not choose a best answer after a week, the status of the question changes to "in-voting", and any user can vote for a best answer until a clear best answer is chosen. Only after this is the question considered "resolved" [8].

Cosine similarity is the most popular algorithm used to find similarity between documents [1]. Cosine similarity converts documents to vectors and finds the cosine of the angle between document vectors. The higher the similarity score, the more the similarity between documents.

The algorithm for cosine similarity is as follows: Let A = ($a_1$, $a_2$, …, an) and B = ($b_1$, b2,…, bn) be two vectors of length n where all the coordinates are positive. The Cosine similarity between vectors A and B is defined as:

$$\frac{A.B}{|A\|B|}$$

Where *A.B* is the inner product of the two vectors *A* and B and |*A*| and |*B*| are the Euclidean norms of *A* and *B.*

The Cosine similarity between two documents corresponds to the correlation between these documents vectors [2]. Cosine similarity values are in the interval [0,1]. Cosine similarity has some disadvantages some of which are:

I) The order in which terms (words) appear in a document is lost in the vector space representation.

II) It does not take into consideration that English words and sentences can be expressed in various ways without losing its meaning or being ambiguous. For example, Cosine similarity will not know that New York City and The Big apple are referring to the same entity, New York City.

## 3. ALGORITHM OVERVIEW

The proposed algorithm searches a dataset of resolved questions from Yahoo! Answers for the most similar question(s) to an unanswered question. The algorithm exploits word synonyms when expanding each unanswered question i.e. rather than search for similar resolved questions to an unanswered question, it expands the unanswered question and searches for similar resolved questions to the unanswered question and its expanded questions.

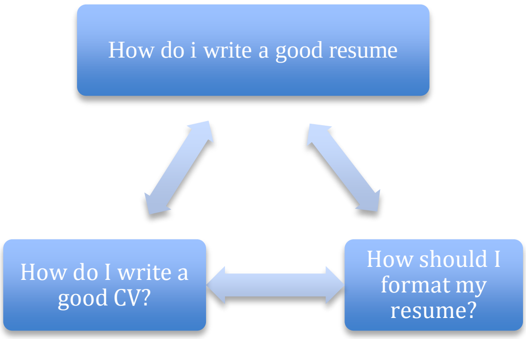
How do i write a good resume
How do I write a good CV?
How should I format my resume?

**Figure 2: shows different ways in which a sentence can be represented.**

This paper is focused on manner questions (how to questions), which are mostly simple and short in length.

Given an unanswered question, the proposed algorithm finds the noun(s) in the unanswered questions and uses different algorithms to find synonyms to these nouns(s). With these synonyms, a question is expanded into multiple related questions as follows:

I) For each synonym of a noun, create a new question by replacing the noun with a synonym.

II) Cosine similarity is used to find and rank similar resolved questions to each of these expanded questions.

III) The algorithm recognizes the NIL case where no word synonym is found.

## 4. PROPOSED ALGORITHM

This section discusses the proposed algorithm, which expands an unanswered question into multiple related questions and looks for the most similar resolved questions to the set of unanswered question and its expanded questions.

### 4.1 Extract Nouns

Most English questions contain nouns and so given an unanswered question, this step extracts all the nouns in question. For example, given the unanswered question " How do I write a good resume?" this step extracts "resume" as the noun in this sentence. WordNet was used to extract the noun(s) in each question.

### 4.2 Noun Synonyms and Approximate String Matching

English words can be expressed in various ways without losing it's meaning by using synonyms and/or abbreviations. For each noun found in the previous step, extract its synonyms and abbreviations using the methods below:

I) Extract each nouns synonyms using WordNet's synonym finder

II) Find word abbreviations. For example CV is an abbreviation for curriculum vitae.

III) Dice coefficient and hamming distance are used to resolve minor spelling errors.

### 4.3 Expand Question (EXPAND)

Given an unanswered question, expand this question into multiple questions by replacing the noun with each found synonym/abbreviation/misspelled word

### 4.4 Resolve Question Similarity

The unanswered question and the expanded questions form a set of questions, which we call EXPAND.

Cosine similarity is used to measure the similarity between each question in EXPAND and the resolved questions. The resolved question with the highest cosine similarity score is selected.

## 5. EXPERIMENTS

This paper focuses on manner questions. Manner questions are usually short in length compared to comprehensive questions and they usually contain at least one noun.

### 5.1 Data Set

A dataset of 300,000 resolved questions from Yahoo! Answers was used for this research. The dataset was provided as part of Yahoo Research Alliance Webscope program. The dataset is from Yahoo Answers 10/25/2007 dump. Each resolved question contains the following:

I) The category and sub-category that was assigned to this question.

II) The subject part of the question with a brief description of the question.

III) The content part where the question is written in more detail.

IV) The best answer

V) Other answers

## 5.2 Experiment on Dataset

The Natural language processing toolkit, NLTK part of speech finder was used to extract nouns from unanswered questions. Different unanswered questions were used for experimentation. One of the question used was “How do I write a good resume?” The noun in this question is “resume” and its synonyms are “curriculum vitae” and “CV”.

WordNet’s synonym finder extracted {sketch, curriculum vitae, sum up} as synonyms of “resume”. Sketch and sum up have no relationship with “resume”.

We tested WordNet’s synonym finder with some nouns and in general, it finds at least one synonym of a noun, albeit it finds some irrelevant words.

Hamming distance and Dice coefficient were used to find similar word(s) to the noun(s). The noun abbreviations were also resolved.

cat>First Aid</cat>

<maincat>Health</maincat>

<subcat>General Health Care</subcat>

</document></vespaadd>

<vespaadd><document type="wisdom">

<uri>126792</uri>

<subject>how do you seperate h2o to make oxygen and hydrogen?</subject> <bestanswer>Electrolysis of water can be achieved in a simple hands-on project, where electricity from a battery is run into a cup of water. Hydrogen gas will be seen to bubble up at one of the immersed battery probes, and oxygen will bubble at the other.2H2O(l) ? 2H2(g) + O2(g)</bestanswer> <nbestanswers><answer_item>Electrolysis of water can be achieved in a simple hands-on project, where electricity from a battery is run into a cup of water. Hydrogen gas will be seen to bubble up at one of the immersed battery probes, and oxygen will bubble at the other. 2H2O(l) ? 2H2(g) + O2(g)</answer_item> <answer_item>electrolysis (a current of electricity) &#xa;just took a test on it last week ;)</answer_item> </nbestanswers>

**Figure 3: Shows the format of the dataset.**

The proposed algorithm expands an unanswered question into multiple questions by creating a new question for each found synonym. For each synonym, the algorithm makes a copy of the unanswered question and replaces the noun with the synonym. The unanswered question and the new questions form a set of questions called “EXPANDED questions”.

The algorithm performs cosine similarity between each of the EXPANDED questions and the resolved questions from Yahoo! Answers. The resolved question with the highest cosine similarity is chosen.

Yahoo! Answers dataset contains 300,000 resolved questions in different categories. For our experiments, unanswered questions from the top 20 categories in Yahoo! Answers website were used to test the proposed algorithm. The precision and recall graph was used to show the performance of both the proposed algorithm and cosine similarity. The precision measures the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. The recall measures the ratio of the number of relevant records retrieved to the total number of relevant records in the dataset. The precision and recall graph in Figure 4 shows that the proposed algorithm despite using only the subject part of unanswered questions extracted more relevant resolved questions than the cosine similarity method used on the subject and content part of the unanswered question.

For example, the question “How do I write a good resume?” is expanded to the following questions:

I) How do I write a good curriculum vitae?

II) How do I write a good sketch?

III) How do I write a good sum up?

The proposed algorithm finds the following relevant results (sorted in descending order of their relevance) from the Yahoo Answers dataset.

I) How do I create a perfect resume?

II) How do I prepare a Pastoral Resume?

III) How can I find instructions for setting up curriculum vitae for myself on the internet?

Cosine similarity found the following results for the Yahoo Answers dataset.

I) How do I create a perfect resume?

II) How do I prepare a Pastoral Resume?

The proposed algorithm found one more relevant resolved question.

Prior to running the experiments, the relevant resolved questions to each unanswered question were marked. The proposed algorithm was used to extract relevant questions from the dataset. This information was used to compute the precision and recall graph.

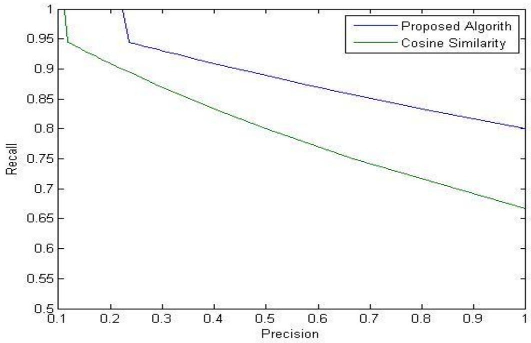

**Figure 4: The Precision and Recall graph shows that when searching for question similarities, incorporating word synonyms in the search returns more relevant results than using cosine similarity only.**

## 6. CONCLUSION AND FUTURE WORK

In order to reduce the number of unanswered questions in question and answering systems, this paper proposed an algorithm to find the most similar question to an unanswered question. The proposed algorithm exploits word synonyms to extract more relevant resolved questions from the dataset. The algorithm performs better than cosine similarity as shown in the precision and recall graph in Figure 4.

In the future we plan to find the best answer to an unanswered question by doing the following:

I) Use the proposed algorithm in this paper to find the relevant questions.

II) Find the answer to one of the relevant questions in the previous step that can be used to satisfactorily answer the unanswered question.

## 7. REFERENCES

[1] Salton, G. and McGill, M, 1987. Introduction to Modern Information Retrieval. McGraw-Hill, NY.

[2] Similarity Measures for Text Document Clustering A. Huang, University of Waikato, Hamilton, New Zealand

[3] Rao, D., McNamee P., and Dredze M. 2010 Entity Linking: Finding Extracted Entities in a Knowledge Base

[4] Gyongyi, Z., Koutrika, G, Pederson, J, and Garcia-Molina, Hector. Questioning Yahoo! Answers.

[5] Lopez, V, Unger, Christina, Cimiano, P, Motta, Enrico. Evaluating Question Answering over Linked Data.

[6] Zhou, T, Si, X, Chang, E, King, I, and Lyu, M. A Data driven approach to Question subjectivity identification in community question answering.

[7] Yih, W, Chang, M, Meek, C, Pastusiak, A. Question Answering Using enhanced Lexical Semantic Models.

[8] Dror, G, Koren, Y, Maarek, Y, Szpektor, I want to Answer, Who has a question? Yahoo! Answers Recommender System.

[9] Shtok, A, Dror, G, Maarek, Y, Szpektor, I. Learning from the Past: Answering New Questions with Pas