Name Variation in Community Question Answering Systems

Anietie Andy
Howard University
anietie.andy@bison.howard.edu

Satoshi Sekine
New York University
sekine@cs.nyu.edu

Mugizi Rwebangira
Howard University
rweba@scs.howard.edu

Mark Dredze
Johns Hopkins University
mdredze@cs.jhu.edu

Abstract

Community question answering systems are forums where users can ask and answer questions in various categories. Examples are Yahoo! Answers, Quora, and Stack Overflow. A common challenge with such systems is that a significant percentage of asked questions are left unanswered. In this paper, we propose an algorithm to reduce the number of unanswered questions in Yahoo! Answers by reusing the answer to the most similar past resolved question to the unanswered question, from the site. Semantically similar questions could be worded differently, thereby making it difficult to find questions that have shared needs. For example, *Who is the best player for the Reds?* and *Who is currently the biggest star at Manchester United?* have a shared need but are worded differently; also, *Reds* and *Manchester United* are used to refer to the soccer team *Manchester United football club*. In this research, we focus on question categories that contain a large number of named entities and entity name variations. We show that in these categories, entity linking can be used to identify relevant past resolved questions with shared needs as a given question by disambiguating named entities and matching these questions based on the disambiguated entities, identified entities, and knowledge base information related to these entities. We evaluated our algorithm on a new dataset constructed from Yahoo! Answers. The dataset contains annotated question pairs, \((Q_{\text{given}}, [Q_{\text{past}}, \text{Answer}])\). We carried out experiments on several question categories and show that an entity-based approach gives good performance when searching for similar questions in entity rich categories.

1 Introduction

In community question answering (CQA) systems, users prefer asking other users questions because (I) their questions are personal and require a direct answer from users with similar experiences or users familiar with the question (II) no single web page can answer their question, and (III) users want to communicate and exchange ideas with other users. One of the challenges with such systems is that some questions are left unanswered because:

- they are short and lack relevant content
- they are not clearly expressed
- they are not appropriately assigned to a user that is able to answer the question

Approximately 15% of incoming English questions in Yahoo! Answers do not receive any answer and leave the user that asked the question (asker) unsatisfied (Shtok et al., 2012). One approach to reducing...
the number of unanswered questions in a CQA is to direct an unanswered question to a user knowledgeable about the question (Dror et al., 2011). Another approach automatically extracts answers from a knowledge base (KB) such as Wikipedia, text passage, or the web (Gyongyi et al., 2007). In certain question categories in Yahoo! Answers, approximately 25% of questions are recurrent (Shtok et al., 2012). A third approach takes advantage of this question recurrence by reusing past resolved questions (PARQ) from within Yahoo! Answers to satisfy unanswered questions. Shtok et al. (2012) used this third approach to satisfy unanswered questions in the Beauty & Style, Health, and Pets question categories by matching new questions to PARQ’s if they had a cosine similarity score above a threshold (0.9); features were then extracted from the new question and PARQ’s to train a classifier. Certain question categories such as Sports have a high occurrence of named entities and entity name variations. For example, a sports team can be referred to by its official name, the name of the city it plays in or by any of several nicknames. Also, the vocabulary in questions in these categories can be diverse and questions are often very short (Klang and Nugues, 2014; Khalid et al., 2008).

The contribution of this paper is to propose an alternative approach to reducing the number of unanswered questions in question categories that contain a large number of entities by taking advantage of the recent successes in entity linking. We now have systems that can disambiguate named entities to a KB. Matching questions and answers based on these disambiguated entities, entities, and KB information related to these entities finds most of the relevant answers to a given question.

We investigate the validity of using an entity-based approach in entity rich categories by first analyzing 150 questions from each of the following categories Beauty & Style, Health, Pets, Sports, Entertainment & Music and Parenting.

<table>
<thead>
<tr>
<th>Question category</th>
<th>Number of questions with named entities or entity variations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beauty &amp; Style</td>
<td>70</td>
</tr>
<tr>
<td>Health</td>
<td>73</td>
</tr>
<tr>
<td>Pets</td>
<td>64</td>
</tr>
<tr>
<td>Sports</td>
<td>130</td>
</tr>
<tr>
<td>Entertainment &amp; Music</td>
<td>135</td>
</tr>
<tr>
<td>Parenting</td>
<td>95</td>
</tr>
</tbody>
</table>

Table 1: Number of questions with named entities or entity variations out of 150 questions from each category

Table 1 shows that more questions in the Sports and Entertainment & Music categories contain named entities and/or entity variations. We annotated 200 question pairs that each exhibit shared needs from the Sports and Entertainment & Music question categories. We observed that 82% of the relevant annotated question pairs contain either the same named entity or a variation of the entity. This percentage could increase on a larger dataset. We also observed that the cosine similarity score of the relevant question pairs varied i.e. the cosine similarity was high in some relevant question pairs and low in others. Hence, we propose to use an entity-based approach in question categories with high entity usage.

2 Related Work and Background

Yahoo! Answers is one of the largest and most popular CQA sites with more than 20 question categories. In Yahoo! Answers, there are two parts to a question: (I) the title - a brief description of the question, and (II) the content - a detailed description of the question (Dror et al., 2011). Posted questions are assigned to predefined categories, such as Pets, Sports and Entertainment & Music and these questions can be
answered by any signed-in user. An asked question remains open for four days, or for less if the asker chose a best answer within this period. If no best answer is chosen by the asker, the task is delegated to the community, which votes for the best answer until a clear winner arises. Only then is the question considered resolved. If a question is not answered while open, it is deleted from the site.

Registered users may answer a limited number of questions each day, depending on their level (Dror et al., 2011). Some of the categories of CQA questions are: factoid, opinion-seeking, recommendation, open-ended, and problem solving questions. Different approaches have been proposed to reduce the number of unanswered questions in Yahoo! Answers. Dror et al. (2011) focuses on matching unanswered questions to users that are presumed to be experts in the question topic, i.e., "routes the right question to the right user". This approach uses a multi-channel recommender system technology for associating an unanswered question with potential answerers that are in an "answering mood". Also in this approach, a wide variety of content and social signals users regularly provide to the CQA system are exploited and organized into channels. Gyongyi et al. (2007) automatically generates answers to questions. In this approach, text passages that may contain the answer to an unanswered question are retrieved and ranked. The passage with the highest rank is selected to answer the unanswered question. Cao et al. (2011) proposes an approach to utilizing category information to enhance the performance of question retrieval. This approach combines the global relevance (the relevance of a query to a category) and the local relevance (the relevance of a query to a question in the category). The intuition behind this approach is that the more related a category is to a query, the more likely it is that the category contains questions relevant to the query. The model ranks a historical question based on an interpolation of two relevance scores: one is a global relevance score between the query and the category containing the historical question, and the other is a local relevance score between the query and the historical question (Cao et al., 2011). Bian et al. (2008) attempts to rank past CQA question-answer pairs in response to factual questions. A supervised learning-to-rank algorithm is used to promote relevant past answers to the input question based on textual properties of the question and the answer, as well as indicators for the answerers quality (Shtok et al., 2012). In Bian et al. (2008), the goal is to detect if a relevant answer exists, and the scope is not limited to factual questions. Carmel et al. (2000) proposes to find past questions that are similar to the target question, based on the hypothesis that answers to similar questions should be relevant to the target question. This approach ranks past questions using both inter-question and question-answer similarity, with response to a newly posed questions (Shtok et al., 2012). Jeon et al. (2005) demonstrates that similar answers are a good indicator of similar questions. Once pairs of similar questions are collected based on their similar answers, they are used to learn a translation model between question titles to overcome the lexical chasm when retrieving similar questions (Jeon et al., 2005). Wang et al. (2011) matches similar questions by assessing the similarity between their syntactic parse tree structure (Shtok et al., 2012). This approach retrieves semantically similar questions and questions with shared needs. Shtok et al. (2012) answers unanswered questions by reusing similar PARQ. This approach searches a dataset of PARQ for similar questions to an unanswered question. The answer to the most similar PARQ is used to answer the unanswered question. This approach relies on the intuition that even if personal and narrow, some questions are recurrent enough to allow for at least a few new questions to be answered by past material. Klang and Nugues (2014) shows that resolving entity disambiguations in question answering systems helps in retrieving relevant answers to a question from documents or passages. Given a question, (Klang and Nugues, 2014) uses a named entity disambiguation module to merge entities in a question answering
Strings that could be linked to a unique identifier are merged and a list of synonyms with the resulting set is created. The candidate answers to a question were ranked based on their frequency i.e. the number of candidate occurrences after merging. This approach shows that a candidate merging step using a named entity linking module produces high precision results. Khalid et al. (2008) investigates the impact of named entity normalization (NEN) on two specific information access tasks: document and passage retrieval for question answering (QA). These tasks consist in finding items in a collection of documents, which contain an answer to a natural language question. In the NEN task, a system identifies a canonical unambiguous referent for names like Bush or Alabama (Khalid et al., 2008). Two entity normalization methods based on Wikipedia in the context of both passage and document retrieval for question answering were evaluated. It was found that normalization methods lead to improvements of early precision, for both document and passage retrieval.

3 Entity Name Variation in CQA systems

Due to the lack of uniformity in CQA users writing styles (Khalid et al., 2008), the lack of content in some questions, and the frequent use of entity name variations in question categories with a large number of entities, it is necessary to use an entity-based approach to find PARQ with shared needs to a given question. In order to retrieve most of the relevant PARQ to a given question with high precision, it is important to identify the named entities and entity variations in the given question and PARQ. For example, Q1 and Q2 below are questions with a shared need referring to Pro MLB umpire and Major League Baseball Umpire respectively.

- **Q1**: How does any one become a Pro MLB umpire?
- **Q2**: How can I become a Major League Baseball Umpire?

The proposed algorithm, ENTITY-ALCHEMY has 2 stages:

3.1 Stage 1

Given the question pair, \((Q_{\text{given}}, [Q_{\text{past}}, Answer])\), where \(Q_{\text{given}}\) represents a given question, \(Q_{\text{past}}\) represents a past resolved question, and \(Answer\) is the answer to \(Q_{\text{past}}\) (Shtok et al., 2012), ENTITY-ALCHEMY identifies named entities in \((Q_{\text{given}}\) and \(Q_{\text{past}}\)) and links these entities to an external KB, using entity linking, to find their name variations and anchor phrases (surface form), textual phrases that potentially link to the entity in the KB (Guo et al., 2013). Using the question-title for retrieval of similar questions in a CQA is of highest effectiveness, while using the question body results in lower Mean Average Precision (MAP) (Shtok et al., 2012). In this stage, we identify named entities and entity variations in the question-title section of \(Q_{\text{given}}\) and \(Q_{\text{past}}\). ENTITY-ALCHEMY selects \(Q_{\text{past}}\) as a candidate similar question to \(Q_{\text{given}}\) if both questions have a common entity, entity name variation, or anchor phrases.

3.2 Stage 2

In stage 2, the algorithm extracts features from a pair of \(Q_{\text{given}}\) and \(Q_{\text{past}}\), selected in stage 1. The extracted features are used to score whether the answer to \(Q_{\text{past}}\) can be used to satisfy the given question, \(Q_{\text{given}}\).

3.2.1 Features

*Entities and KB information*: We collect the following statistics from \((Q_{\text{given}}\) and \(Q_{\text{past}})\): number of common entities, number of common entity variations, number of common anchor phrases, number of common words or phrases.
**Lexical Analysis:** We classify words in \((Q_{\text{given}} \text{ and } Q_{\text{past}})\) into their parts-of-speech and extract the number of matching nouns, verbs, and adjectives, if they exist.

**Cosine similarity:** Cosine similarity is popularly used to show the similarity between documents (Salton and McGill, 1986). We calculate the cosine similarity of the "title" and "title + content" of \((Q_{\text{given}} \text{ and } Q_{\text{past}})\).

**Dice coefficient:** Misspelled words are common in CQA systems. We use dice coefficient to calculate the string similarity score between identified entities in \((Q_{\text{given}} \text{ and } Q_{\text{past}})\).

**Word2vec feature:** Mikolov et al. (2013) introduced an efficient implementation of the continuous bag-of-words and skip-gram techniques that can be used for learning high-quality word vectors from huge datasets with billions of words and with millions of words in its vocabulary called word2vec (Mikolov et al., 2013). We trained a word2vec model with a Wikipedia dump and 200 question pairs from Yahoo! Answers.

### 3.2.2 Classifier model

For learning, we used SVM with a polynomial kernel as implemented by Weka machine learning workbench (Hall et al., 2009). The default SVM parameters were used.

### 4 Experiments

#### 4.1 Experimental Setup

For this research we used a repository of PARQ from Yahoo! Answers. Since we are interested in finding PARQ with answers that can satisfy a given question, we selected the best answers for each question in the *Sports* and *Entertainment & Music* question categories. We selected these question categories because of the high recurrence of questions and the high occurrence of named entities and named entity variations in these question categories.

#### 4.2 Data Construction and Labeling

The dataset used to train and evaluate our system contains question pairs, \((Q_{\text{given}}, [Q_{\text{past}}, \text{Answer}])\), where \(Q_{\text{given}}, Q_{\text{past}},\) and \(\text{Answer}\) belong to the Yahoo! Answers repository. Each question pair was associated with a label, described below:

- **Potential answer:** given a question pair, \((Q_{\text{given}}, [Q_{\text{past}}, \text{Answer}])\), \(\text{Answer}\) is a "potential answer" if it can be used to satisfy \(Q_{\text{given}}\).
- **Similar question:** \(Q_{\text{past}}\) is similar to \(Q_{\text{given}}\) if they both refer to the same topic\(^1\), but the answer to \(Q_{\text{past}}\) cannot be used to satisfy \(Q_{\text{given}}\).
- **Related question:** \(Q_{\text{past}}\) is related to \(Q_{\text{given}}\) if it contains a common entity as \(Q_{\text{given}}\), but refers to a different topic from \(Q_{\text{given}}\).

We sampled 1500 resolved questions from the *Sports* and *Entertainment & Music* question categories (750 from each question category) and observed that approximately 20% and 17% respectively of the sampled questions were recurring. To generate the given question and PARQ pair, \((Q_{\text{given}}, [Q_{\text{past}}, \text{Answer}])\), we selected 3000 and 5000 PARQ from the *Sports* and *Entertainment & Music* question categories respectively from the language data section of Yahoo labs Webscope\textsuperscript{TM} dataset, and Yahoo!

\(^1\) A topic is an activity or event along with all directly related events and activities. A question is on topic when it discusses events and activities that are directly connected to the topic’s seminal event.
Answers dataset (Chang et al., 2008). Given a question from the selected dataset of PARQ, we selected a candidate similar question in the selected dataset if it had a common named entity, entity variation or anchor phrase as the given question. We had three independent reviewers label the question pairs as either a potential answer, similar question, or related question. We selected a question pair if at least two of the reviewers agreed on the question pair label. We annotated 500 question pairs from the Sports and Entertainment & Music question categories. Table 2 shows the number of question pairs and their labels in each of the question categories. In each of the question categories, we calculated the reviewer agreements by using Fleiss’ kappa. Table 3 shows the calculated kappa values. This dataset will be provided to the research community.

We used an entity linking tool, AlchemyAPI (Turian, 2013) to extract named entities, named entity disambiguations, and anchor phrases from a given question and a PARQ. AlchemyAPI extracts anchor phrases from the following KB’s, dbpedia and freebase. In our experiments we split our dataset by using 66% for training and 34% for testing. We conducted two baseline experiments on our dataset using SVM described in Section 3.2.1.

<table>
<thead>
<tr>
<th>Category</th>
<th>Sports</th>
<th>Entertainment &amp; music</th>
</tr>
</thead>
<tbody>
<tr>
<td>Potential answer</td>
<td>130</td>
<td>141</td>
</tr>
<tr>
<td>Similar question</td>
<td>64</td>
<td>40</td>
</tr>
<tr>
<td>Related question</td>
<td>65</td>
<td>60</td>
</tr>
</tbody>
</table>

Table 2: Number of question pairs in each question pair category

<table>
<thead>
<tr>
<th>Question Categories</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sports</td>
<td>0.579</td>
</tr>
<tr>
<td>Entertainment &amp; Music</td>
<td>0.55</td>
</tr>
</tbody>
</table>

Table 3: Fleiss’ Kappa calculation in each question category

<table>
<thead>
<tr>
<th>Named Entity</th>
<th>Entity Name Variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>English Premier League</td>
<td>EPL, premier league</td>
</tr>
<tr>
<td>New York City marathon</td>
<td>NYC marathon</td>
</tr>
<tr>
<td>Jonas brothers</td>
<td>Jonas bros</td>
</tr>
<tr>
<td>Manchester United</td>
<td>Man u, munited</td>
</tr>
</tbody>
</table>

Table 4: Some named entities and their name variations in our dataset

Below are examples of question pairs in each question pair category:

**Question pair 1**
*How do you get on Oprah?*

*<potential answer>*

*How do I get on the Oprah Winfrey show?*

**Question pair 2**
*how do i win to Germany to watch the FIFA WORLD CUP?*

*<similar question>*

*how do i get tickets for Fifa Worldcup 2006 in Germany ?*

---

2 Fleiss kappa assesses the reliability of the agreement between the raters when assigning labels to the question pairs.
Question pair 3

How can I get on the Jay Leno show?

<related question>

how do i get salmas hayek interview with jay leno on march 3 2006?

4.3 Evaluation Metric

We measure the **precision**, **recall**, and **accuracy** of the proposed algorithm.

**Precision:** the fraction of returned answers that are correct i.e. potential answers.

**Recall:** the fraction of the labelled potential answer question pairs that where returned by the system.

**Accuracy:** the overall fraction of potential answer question pairs classified correctly.

4.4 Results

The first baseline, **ENT**, uses AlchemyAPI to extract named entities from a question pair. The extracted entities are not disambiguated to a KB and anchor phrases from the named entities KB entries are not extracted. In this baseline, we aim to find the most similar PARQ with common entities as a given question. **ENT** has two stages: In stage 1, given a question pair, \( \langle Q_{\text{given}}, [Q_{\text{past}}, \text{Answer}] \rangle \), we select \( Q_{\text{past}} \) if it contains a common named entity as \( Q_{\text{given}} \). In stage two, we extract the features described in section 3.2.1 from the question pair.

Yahoo! Answers is an informal forum, hence, there is a high prevalence of misspelled words. The second baseline, **ENT-VARIANT** aims to find the most similar PARQ with common entities and minor entity spelling errors as a given question. **ENT-VARIANT** uses AlchemyAPI to identify the named entities in each question in a question pair. **ENT-VARIANT** has two stages. In stage 1, given a question pair, \( \langle Q_{\text{given}}, [Q_{\text{past}}, \text{Answer}] \rangle \), we select \( Q_{\text{past}} \) if it contains a common named entity as \( Q_{\text{given}} \). Also, dice coefficient is used to compare the identified named entities in \( Q_{\text{given}} \) and \( Q_{\text{past}} \). This comparison helps resolve minor spelling errors in the question pair. In our experiments, two named entities in \( Q_{\text{given}} \) and \( Q_{\text{past}} \) respectively, are considered a variation with minor spelling errors if they have a dice coefficient > 0.75. In stage 2 of **ENT-VARIANT**, we extract the features described in section 3.2.1 from the question pair.

**ENTITY-ALCHEMY** performed better than both baselines as shown in Table 5. This shows that identifying named entities, disambiguated entities to a KB and, extracting anchor phrases from the identified named entities KB entries finds more relevant PARQ to a given question.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENT</td>
<td>66%</td>
<td>43.05%</td>
</tr>
<tr>
<td>ENT-VARIANT</td>
<td>67.1%</td>
<td>44.5%</td>
</tr>
<tr>
<td>ENTITY-ALCHEMY</td>
<td>71%</td>
<td>55.15%</td>
</tr>
</tbody>
</table>

Table 5: Precision and recall of **ENTITY-ALCHEMY** and two baselines

For each of the question categories we measured the accuracy, as defined in section 4.3, of the baseline algorithms and **ENTITY-ALCHEMY**. Table 6 presents the accuracy of **ENTITY-ALCHEMY** and the two baselines on the **Sports** and **Entertainment & Music** question categories. Most of the entity variations in our dataset are not minor spelling errors, hence the second baseline, **ENT-VARIANT** did not perform a lot better than the first baseline, **ENT**.
We tested ENTITY-ALCHEMY on a question category, Parenting, which contains few named entities and entity name variations to see how well it will perform. We extracted 500 questions from the Parenting question category of Yahoo! Answers and selected question pairs from this extracted dataset by applying stage 1 of ENTITY-ALCHEMY. Table 7 shows the results of ENTITY-ALCHEMY and cosine similarity on the Sports and Parenting question categories. ENTITY-ALCHEMY identified 51% of similar questions that exhibited shared needs. We also selected question pairs by conducting cosine similarity and our experiments showed that in this question category, a cosine similarity > 0.5 identified 87% of similar questions that exhibited a shared need. We applied cosine similarity to the Sports question category by selecting question pairs that exhibit shared needs from 500 extracted questions from the Sports category. Our experiments show that ENTITY-ALCHEMY identified 83% of the similar question pairs that exhibited a shared need and a cosine similarity > 0.5 identified 49% of the similar questions. Hence in question categories with less entity and entity variation usage, a non-entity-based approach such as cosine similarity should be used to find similar questions with shared needs to a given question. Also, in entity rich question categories, an entity-based approach should be used when searching for questions with shared needs.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Sports</th>
<th>Parenting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cosine Similarity</td>
<td>49%</td>
<td>87%</td>
</tr>
<tr>
<td>ENTITY-ALCHEMY</td>
<td>83%</td>
<td>51%</td>
</tr>
</tbody>
</table>

Table 7: Comparing cosine similarity and entity-based approaches in the Sports and Parenting question categories

5 Conclusion
In this paper, we proposed an algorithm, ENTITY-ALCHEMY, to reduce the number of unanswered questions in question categories with high entity usage. We evaluated our algorithm on a CQA dataset with a lot of entities and entity variations and our algorithm performed better than two baselines.

In conclusion, reusing PARQ is an effective method for reducing the number of unanswered questions in a CQA system. This paper showed that in question categories with a lot of named entities and entity name variations, using KB information and applying entity linking to identify and disambiguate named entities finds most of the similar PARQ to a given question.

6 Future Work
In the future, we would research time-sensitive questions especially common in Sports categories. Also, we would implement an algorithm that can find similar questions to a given question regardless of the question category.

Acknowledgements
This work was supported by the Center for Science of Information (CSOI), an NSF Science and Technology Center, under grant agreement CCF-0939370.
References


Dror, Gideon and Maarek, Yoelle and Szpektor, Idan 2011. I want to answer; who has a question?: Yahoo! answers recommender system Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining 1109–1117

Rao, Delip and McNamee, Paul and Dredze, Mark 2013. Entity linking: Finding extracted entities in a knowledge base Springer Multi-source, multilingual information extraction and summarization 93–115


Voorhees, Ellen M and Tice, DM 2000. Overview of the TREC-9 Question Answering Track Proceedings of TREC

Strzalkowski, Tomek and Harabagiu, Sandra 2006. Advances in open domain question answering, volume 32 Springer Science & Business Media


Chang, Ming-Wei and Ratinov, Lev-Arie and Roth, Dan and Srikrumar, Vivek 2008. Importance of Semantic Representation: Dataless Classification Proceedings of AAAI 830–835


Guo, Stephen and Chang, Ming-Wei and Kiciman, Emre 2013. To Link or Not to Link? A Study on End-to-End Tweet Entity Linking HLT-NAACL 1020–1030

Huang, Anna 2008. Similarity measures for text document clustering Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand 49–56

Mikolov, Tomas and Sutskever, Ilya and Chen, Kai and Corrado, Greg S and Dean, Jeff 2013. Distributed representations of words and phrases and their compositionality Advances in neural information processing systems 3111–3119


