Exploiting Named Entity Synonyms in Question Answering Systems

Anietie Andy
Computer Science Department
Howard University
202-486-4095
anietie.andy@bison.howard.edu

Mugizi Rwebangira
Computer Science Department
Howard University
rweba@scs.howard.edu

ABSTRACT
Community Question answering (CQA) systems have increasingly become popular among internet users. CQA's such as Yahoo! Answers have a large repository of resolved questions i.e. questions that have been satisfactorily answered. One of the challenges with these systems is that some questions are left unanswered thereby leaving the user unsatisfied. Papers have proposed algorithms to reduce the number of unanswered questions. Some of the proposed algorithms work as follows: (i) direct the unanswered question to CQA users that can potentially answer the question and (ii) use answers to past resolved questions that are similar to the given question. This paper explores an approach that extracts the word synonyms of named entities in a given question and searches the dataset of past resolved questions for similar resolved questions to the given question. The answer to the most similar resolved question is used to satisfy the given question. This paper proposes an algorithm to improve the CQA user experience by using the answer to the most similar past resolved question to satisfy a given question. In cases where the answer to the most similar question cannot satisfy the given question, the proposed algorithm recommends the answer to a related question. Although the answer to the related question will probably not satisfy the given question, it will engage the CQA user and perhaps provide a clue to answer the given question.

Keywords
Community question answering, Information extraction

1. INTRODUCTION

Yahoo! Answers provides a forum where users can ask questions – some of which are personal, that require a direct answer from other users. Also, in Yahoo! Answers, users can engage each other by exchanging ideas / opinions about topics and questions of interest.

Yahoo! Answers has at least 20 question categories and questions in these categories cover various topics and interests. Some of the question categories in Yahoo! Answers such as *Sports* contain more named entities than some other question categories. So, in order to extract most of the similar past resolved questions to a given question, it is important to first identify the question category. If the question category contains a large number of named entities, the method used to search for similar past resolved questions should take into consideration the synonyms of the named entities in the given question and past resolved questions.

A question in Yahoo! Answers has two parts: The *Subject* - a brief description of the question and the *Content* - a more detailed description of the question. A Yahoo! Answers user can ask a question in any question category. Other CQA users respond to the asked question with their answers or ideas. One of the benefits of a CQA is the social aspect of users communicating by either asking or responding to questions or ideas. A user asking a question (asker) can engage users answering her question (answerers). The asker selects the best answer from among the suggested answers that satisfies her question.

One approach to reducing the number of unanswered questions in Yahoo! Answers is to "route the right question to the right user" [2]. This approach uses a multi-channel recommender system technology for associating questions with potential answerers of Yahoo! Answers that are in an "answering mood" [2]. When in an "answering mood", answerers mainly skim through long and dynamic lists of open questions. This approach exploits a wide variety of content and social signals users regularly provide to the CQA system and organizes them into channels. The content signals relate mostly to the text and categories of questions and associated answers; social signals capture the various user interactions with questions such as asking, answering, and voting. Because answerers are not employees of the CQA system, there is no guarantee that a question will be satisfactorily answered.

15% of incoming English questions in Yahoo! Answers do not receive any answer and leave the asker unsatisfied [1]. However, English named entities can have multiple representations e.g. "*Oprah*" and "*Oprah Winfrey show*"; therefore it is possible to express the same idea in multiple ways. For example:

Q1: How do you get on Oprah?

Q2: How do I get on the Oprah Winfrey show?

Q1 and *Q2* above are two similar questions referring to *Oprah* and the *Oprah Winfrey show* respectively.

Given *Q1* above as a given question and *Q2* as a past resolved question. The answer to *Q2* can satisfy *Q1*. *Oprah* in *Q1* is referring to *the Oprah Winfrey show* and not *Oprah Winfrey*, the host of the *Oprah Winfrey show*.

In this paper, we propose to apply a module to identify synonyms of named entities in questions in CQA's.

The proposed algorithm contributes the following:

1. Identification of synonyms of named entities in a community question answering system.
2. Recommend answers to the most similar past resolved question to a given question.
3. If the given question does not have a corresponding similar past resolved question, the algorithm will recommend the most related question if it exists.

2. Synonyms in community question answering systems

The proposed algorithm will have 2 steps. In the first step, named entities will be identified and their synonyms will be extracted. Also in this step, the algorithm will select candidate past resolved questions that contains either the named entity or at least one of

the identified synonyms in the given question and has a cosine similarity greater than a threshold.

In the second step, features will be extracted from the given question, past resolved question and its corresponding answer and a machine learning classifier will be trained to select and recommend the answer to the most similar past resolved question to the given question.

In this paper, we define a named entity as a non-ambiguous, terminal page in Wikipedia (i.e. a Wikipedia page that is not a category, disambiguation, list or redirect page) [3].

2.1 Step 1 of proposed algorithm

Given a question, our algorithm identifies named entities such as *people, locations, and organizations* in the question, if they exist. For example, given the question *Q3* below:

Q3: *how do i get to watch the uefa champions league here in the USA?*

The proposed algorithm identifies *uefa* as an *organization* named entity.

Many named entities in English have synonyms and acronyms. In this step, the proposed algorithm also extracts the synonyms of the identified named entity.

As described in section 1, questions in Yahoo! Answers are made up of two sections: subject and content. [4, 1] show that using the question title for retrieval of similar questions is of highest effectiveness, while using the question body (content) resulted in lower MAP. We measure the similarity between the title of the given question and the past resolved question in step 1 of the algorithm by selecting candidate past resolved questions that have a cosine similarity greater than a threshold and also contain either the identified named entity or a synonym / abbreviation of the identified named entities in the given question.

2.2 Step 2 of proposed algorithm

In step 1, some of the similar past resolved questions to a given question are selected. This considerably reduces the number of similar past resolved questions that the proposed algorithm will have to process.

In step 2, we extract features from the given question, the candidate questions and answers extracted from step 1 above and train a machine learning classifier to select the answer to the most similar past resolved question to the given question. In this paper we will use the support vector machine (SVM) classifier. Below is a description of the extracted features:

2.2.1 Features

2.2.1.1 Textual features:

We extract the Part-of-Speech tag and tf-idf from the given question and candidate past resolved questions and answers. We count the number of similar nouns, verbs, adverbs, adjectives, and stop words that the given question and candidate past resolved question have in common. We count the number of question marks, punctuations, and lengths of the given question and the candidate similar past resolved question; these features measure the complexity of the question.

2.2.1.2 Cosine similarity:

We calculate the cosine similarity of the subject section and the subject + content sections of the given question and past resolved question.

2.2.1.3 Question topic identification

Similar to [1], for each category of Yahoo! Answers we learn the latent dirichlet allocation (LDA) topics [5] from the dataset of past resolved questions in the category. We then infer the distribution of topics from the given question, similar past resolved question, and answer to past resolved answer. From these distributions, we generate topic quality features for each entity by measuring the entropy of the topic distribution and extracting the probability of the most probably topic [1].

3. Experimental setup

Yahoo! Research makes several of their datasets available to university-affiliated researchers through their Webscope program. In this research, we will use a dataset from Yahoo! Answers manner questions from the language data section of Yahoo labs. This dataset contains 300,000 resolved questions in various question categories. Each resolved question contains the following:

1. The category and sub-category that was assigned to the question.
2. The subject part of the question with a brief description of the question.
3. The content part where the question is written in more detail.
4. The best answer that satisfied the question. The asker selects the best answer.
5. Other answers provided by answerers but were not selected as the best answer.

In this research, we use questions from the *Sports* category of Yahoo! Answers. We chose this category because of (I) the high recurrence of question in this category (II) most questions in the sports category contain at least one named entity, and (III) the high occurrence of named entity variations in questions in these categories.

We sampled 250 questions from the sports category of Yahoo! Answers and approximately 70% of the questions in this category contained either a named entity or a variation of a named entity. Also, we sample 50 named entities from questions in the sports category of Yahoo! Answers and 40% of these named entities had at least one reference to a name variation.

4. Conclusion

In conclusion, from the analysis of the dataset (sports category of Yahoo! Answers) discussed in section 3 above, it is evident that applying the proposed algorithm to the sports question category of Yahoo! Answers will reduce the number of unanswered questions thereby improving the experience of CQA users.

5. REFERENCES

[1] Shtok, A., Dror, G., Maarek, Y., & Szpektor, I. (2012, April). Learning from the past: answering new questions with past answers. In *Proceedings of the 21st international conference on World Wide Web* (pp. 759-768). ACM.

[2] Dror, G., Koren, Y., Maarek, Y., & Szpektor, I. (2011, August). I want to answer; who has a question?: Yahoo! answers recommender system. In*Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1109-1117). ACM.

[3] Guo, S., Chang, M. W., & Kiciman, E. (2013). To Link or Not to Link? A Study on End-to-End Tweet Entity Linking. In *HLT-NAACL* (pp. 1020-1030).

[4] Jeon, J., Croft, W. B., & Lee, J. H. (2005, October). Finding similar questions in large question and answer archives. In *Proceedings of the 14th ACM international conference on Information and knowledge management* (pp. 84-90). ACM.

[5] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, *3*, 993-1022.

[6] Chang, M. W., Ratinov, L. A., Roth, D., & Srikumar, V. (2008, July). Importance of Semantic Representation: Dataless Classification. In *AAAI*(pp. 830-835).

[7] Andres Corrada-Emmanuel, W Bruce Croft, and Vanessa Murdock. Answer passage retrieval for question answering. Center Intell. Inf. Retrieval, Univ. Massachusetts, Amherst, MA, Tech. Rep.fOnlineg. Available:http://ciir. cs. umass. edu/pubfiles/ir-283. pdf, 2003.

[8] Gyongyi, Z., Koutrika, G., Pedersen, J., & Garcia-Molina, H. (2007). Questioning yahoo. *Answers: Stanford InfoLab*.

[9] Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen F¨urstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. Robust disambiguation of named entities in text. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 782–792. Associationfor Computational Linguistics, 2011.

[10] Yun Zhou and W Bruce Croft. Query performance predictionin web search environments. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pages 543–550. ACM, 2007.

[11] Xue, X., Jeon, J., & Croft, W. B. (2008, July). Retrieval models for question and answer archives. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 475-482). ACM.

[12] Voorhees, E. M., & Tice, D. M. (2000, November). Overview of the TREC-9 Question Answering Track. In *TREC*.

[13] Tellex, S., Katz, B., Lin, J., Fernandes, A., & Marton, G. (2003, July). Quantitative evaluation of passage retrieval algorithms for question answering. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval* (pp. 41-47). ACM.

[14] Marcus Klang and Pierre Nugues. Named entity disambiguation in a question answering system. In TheFifth Swedish Language Technology Conference (SLTC2014), 2014.

[15] Mahboob Alam Khalid, Valentin Jijkoun, and MaartenDe Rijke. The impact of named entity normalizationon information retrieval for question answering.In Advances in Information Retrieval, pages 705–710. Springer, 2008.