Towards Automatic Question Answering over Social Media by Learning Question Equivalence Patterns

Tianyong Hao[1]
City University of Hong Kong
81 Tat Chee Avenue
Kowloon, Hong Kong SAR
haotianyong@gmail.com

Wenyin Liu
City University of Hong Kong
81 Tat Chee Avenue
Kowloon, Hong Kong SAR
csliuwy@cityu.edu.hk

Eugene Agichtein
Emory University
201 Dowman Drive
Atlanta, Georgia 30322 USA
eugene@mathcs.emory.edu

Abstract

Many questions submitted to Collaborative Question Answering (CQA) sites have been answered before. We propose an approach to automatically generating an answer to such questions based on automatically learning to identify "equivalent" questions. Our main contribution is ***an unsupervised method for automatically learning question equivalence patterns from CQA archive data***. These patterns can be used to match new questions to their equivalents that have been answered before, and thereby help suggest answers automatically. We experimented with our method approach over a large collection of more than 200,000 real questions drawn from the Yahoo! Answers archive, automatically acquiring over 300 groups of question equivalence patterns. These patterns allow our method to obtain over 66% precision on automatically suggesting answers to new questions, significantly outperforming conventional baseline approaches to question matching.

1 Introduction

Social media in general exhibit a rich variety of information sources. Question answering (QA) has been particularly amenable to social media, as it allows a potentially more effective alternative to web search by directly connecting users with the information needs to users willing to share the information directly (Bian, 2008). One of the useful by-products of this process is the resulting large archives of data – which in turn could be good sources of information for automatic question answering. Yahoo! Answers, as a collaborative QA system (CQA), has acquired an archive of more than 40 Million Questions and 500 Million answers, as of 2008 estimates.

The main premise of this paper is that there are many questions that are syntactically different while semantically similar. The key problem is how to identify such question groups. Our method is based on the key observation that *when the best non-trivial answers chosen by asker in the same domain are exactly the same, the corresponding questions are semantically similar*. Based on this observation, we propose answering new method for learning question equivalence patterns from CQA archives. First, we retrieve "equivalent" question groups from a large dataset by grouping them by the text of the best answers (as chosen by the askers). The equivalence patterns are then generated by learning common syntactic and lexical patterns for each group. To avoid generating patterns from questions that were grouped together by chance, we estimate the group's *topic diversity* to filter the candidate patterns. These equivalence patterns are then compared against newly submitted questions. In case of a match, the new question can be answered by proposing the "best" answer from a previously answered equivalent question.

We performed large-scale experiments over a more than 200,000 questions from Yahoo! Answers. Our method generated over 900 equivalence patterns in 339 groups and allows to correctly suggest an answer to a new question, roughly 70% of the time – outperforming conventional similarity-based baselines for answer suggestion.

Moreover, for the newly submitted questions, our method can identify equivalent questions and generate equivalent patterns incrementally, which can greatly improve the feasibility of our method.

2 Learning Equivalence Patterns

While most questions that share exactly the same "best" answer are indeed semantically equivalent, some may share the same answer by chance. To

[1] Work done while visiting Emory University

Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media, pages 9–10,
Los Angeles, California, June 2010. ©2010 Association for Computational Linguistics

filter out such cases, we propose an estimate of Topical Diversity (TD), calculated based on the shared topics for all pairs of questions in the group. If the diversity is larger than a threshold, the questions in this group are considered *not* equivalent, and no patterns are generated. To calculate this measure, we consider as topics the "notional words" (NW) in the question, which are the head nouns and the heads of verb phrases recognized by the OpenNLP parser. Using these words as "topics", *TD* for a group of questions *G* is calculated as:

$$TD(G) = \frac{2}{n(n-1)} \times \sum_{i=1}^{n-1} \sum_{j=2}^{n} (1 - \frac{Q_i \mathbf{I} Q_j}{Q_i \mathbf{U} Q_j}) \quad (i < j)$$

where Q_i and Q_j are the notional words in each question in within group *G* with n questions total.

Based on the question groups, we can generate equivalence patterns to extend the matching coverage – thus retrieving similar questions with different syntactic structure. OpenNLP is used to generate the basic syntactic structures by phrase chunking. After that, only the chunks which contain NWs are analyzed to acquire the phrase labels as the syntactic pattern. Table 1 shows an example of a generated pattern.

Question: What was the first book you discovered that made you think reading wasn't a complete waste of time? **Pattern**: [NP]-[VP]-[NP]-[NP]-[VP]-[VP]-[NP]-[VP]-… **NW**: **(Disjoint**: read waste time) (**Shared**: book think)
Question: What book do you think everyone should have at home? **Pattern**: [NP]-[NP]-[VP]-[NP]-[VP]-[PP]-[NP] **NW**: **(Disjoint**: do everyone have home) (**Shared**: book think)

Table 1. A group of equivalence patterns

3 Experimental Evaluation

Our dataset is 216,563 questions and 2,044,296 answers crawled from Yahoo! Answers. From this we acquired 833 groups of similar questions distributed in 65 categories. After filtering by topical diversity, 339 groups remain to generate equivalence patterns. These groups contain 979 questions, with, 2.89 questions per group on average.

After that, we split our data into 413 questions for training (200 groups) and 566 questions, with randomly selected an additional 10,000 questions, for testing (the remainder) to compare three variants of our system Equivalence patterns only (EP), Notional words only (NW), and the weighted combination (EP+NW). To match question, both equivalence patterns and notional words are used with different weights. The weight of pattern, disjoint NW and shared NW are 0.7, 0.4 and 0.6 after parameter training. We then compare the variants and results are reported in Table 2, showing that EP+NW achieves the highest performance.

	Recall	Precision	F1 score
EP	0.811	0.385	0.522
NW	0.378	0.559	0.451
EP+NW	0.726	0.663	0.693

Table 2. Performance comparison of three variants

Using EP+NW as our best method, we now compare it to traditional similarity-based methods on whole question set. TF*IDF-based vector space model (TFIDF), and a more highly tuned Cosine model (that only keeps the same "notional words" filtered by phrase chunking) are used as baselines. Figure 3 reports the results, which indicate that EP+NW, outperforms both Cosine and TFIDF methods on all metrics.

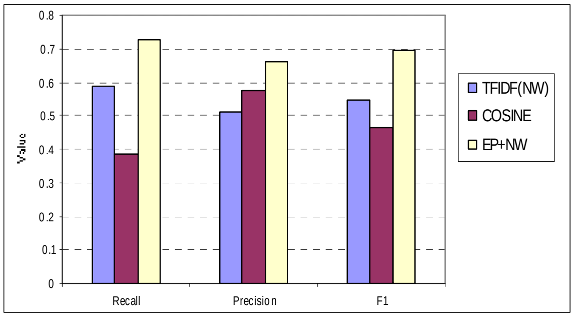

Figure 3. Performance of EP+NW vs. baselines

Our work expands on previous significant efforts on CQA retrieval (e.g., Bian et al., Jeon et al., Kosseim et al.). Our contribution is a new unsupervised and effective method for learning question equivalence patterns that exploits the structure of the collaborative question answering archives – an important part of social media.

4 References

Bian, J., Liu, Y., Agichtein, E., and Zha, H. 2008. *Finding the right facts in the crowd: factoid question answering over social media*. WWW.

Jeon, J., Croft, B.W. and Lee, J.H. 2005. *Finding similar questions in large question and answer archives Export Find Similar*. CIKM.

Kosseim, L. and Yousefi, J. 2008.*Improving the performance of question answering with semantically equivalent answer patterns*, Journal of Data & Knowledge Engineering.