Educational Question Answering based on Social Media Content

Iryna GUREVYCH [1], Delphine BERNHARD, Kateryna IGNATOVA and Cigdem TOPRAK

Ubiquitous Knowledge Processing (UKP) Lab, Computer Science Department, Technical University of Darmstadt, Germany

Abstract. We analyze the requirements for an educational Question Answering (QA) system operating on social media content. As a result, we identify a set of advanced natural language processing (NLP) technologies to address the challenges in educational QA. We conducted an inter-annotator agreement study on subjective question classification in the Yahoo!Answers social Q&A site and propose a simple, but effective approach to automatically identify subjective questions. We also developed a two-stage QA architecture for answering learners' questions. In the first step, we aim at re-using human answers to already answered questions by employing question paraphrase identification [1]. In the second step, we apply information retrieval techniques to perform answer retrieval from social media content. We show that elaborate techniques for question preprocessing are crucial.

Keywords. question answering, social media, question subjectivity

1. Introduction

In recent years, the amount of digital textual information has been constantly increasing, leading to the well-known *information overload* problem. While this problem is especially acute for learners, conventional search engines are often ill-suited to address learners' complex information needs. We believe that Question Answering (QA) represents a more appropriate Natural Language Processing (NLP) technology in educational contexts, both to reduce the learners' information overload and the instructors' work overload. On the one hand, learners have to deal with a growing amount of learning and community-based material in which to look for relevant information. On the other hand, instructors are overwhelmed with students' questions asked via forums or emails. These challenges should be addressed by an educational QA system which could automatically answer a significant part of the students' questions. Educational QA would thus constitute a significant technological asset for independent and technology-enhanced learning.

QA systems actually share some interesting characteristics with other learning technologies. They provide a means for learners to obtain answers to their questions, just as forums and chats. However, QA systems are not dependent on human responses and thus cater for timely responses. They are also related to Intelligent Tutoring sytems, though

[1]Corresponding Author: Iryna Gurevych, UKP Lab, Technical University of Darmstadt, Hochschulstr. 10, 64289 Darmstadt, Germany; E-mail: gurevych@tk.informatik.tu-darmstadt.de

less complex since most of the time QA systems do not support dialogues.[2] In contrast to most ITS systems also, QA is not limited to a single domain.

Traditional QA systems, such as the system described by Hovy et al. [2], can be decomposed in several modules. Questions are first processed to identify the question class and the expected answer type. Then, an information retrieval module identifies relevant documents. In the third step, relevant documents are split into topical segments and candidate answers are selected. Eventually, the best answers are identified and ranked. Unfortunately, state of the art QA systems suffer from several shortcomings which make them ill suited for educational uses. First, they are usually targeted at factoid questions, while learners' questions are usually long and open-ended and cannot be answered by a single sentence [3,4]. Second, they expect perfectly formulated questions [5], while question asking practice, as displayed in social Q&A sites or query logs, shows that real user questions are often ill-formulated and contain grammatical and spelling errors. Third, the quality of an answer has to be verified by an answer processing module.

The datasets where the answers will be sought for also constitute an important aspect of QA systems, since they directly influence the performance and coverage of the QA system. We propose to use social media content for answer searching. Social media and Web 2.0 tools have recently entered the classroom and have been put to use for different pedagogical objectives: blogs to gather student comments on a specific assignment or topic, wikis for collaborative writing projects etc. This has led to the production of huge amounts of user generated content, which contains a lot of educationally relevant information and which can be employed in educational applications and especially educational QA. Since this content is of variable quality, the answers extracted from user generated discourse such as wikis or forums have to be assessed before they are displayed to the user.

In this article, we analyse the technological requirements for an educational QA system designed to support the learners while searching for relevant information in social media content. We discuss the corresponding system architecture (Section 2) and present experimental work targeted at its several components. Firstly, we present a corpus-based study of subjective questions and an effective lexicon-based approach for subjective question identification (Section 3). Secondly, we apply information retrieval techniques to answer real user questions from social Q&A sites and show the importance of question analysis. In Section 5, we summarize the requirements for an educational QA system operating on social media content as well as the main findings of this paper. We also outline further research needed to enable highly usable educational QA systems.

2. Architecture of an Educational Question Answering System

An educational QA system entails a set of additional challenges as compared to conventional QA systems. The system architecture (Figure 1) gives an overview of how we propose to meet them. In a previous study [6], we found that a large proportion of questions in social Q&A sites is ill-formed. Very often, learners have difficulties in formulating a good question. Therefore, each question should be assessed for its quality and the question paraphrasing component is utilized to generate a high-quality question from a low-quality question. In the next step, the question type should be identified to adjust

[2] There has however been some recent advances towards *interactive QA*.

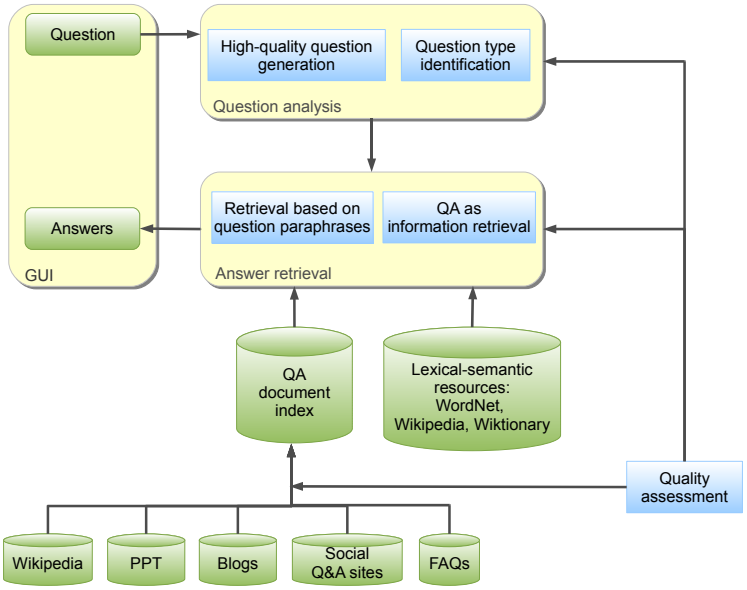

Figure 1. Architecture of an Educational Question Answering System.

the answer processing of a QA system according to the type of the question. In previous work [7], we adopted the Graesser question classification scheme to analyze question types in social Q&A sites. In the present paper, we focus on the manual and automatic classification of subjective questions, see Section 3. Subjectivity analysis of answers and question type classification for non-subjective questions are left to future work.

The answer retrieval component of the system is divided in two principal steps. The first step aims at finding already answered questions, which are paraphrases of the question at hand. The main NLP technology utilized in this step is question paraphrase identification [1]. If no exact question paraphrase is found, similar questions generated via question paraphrasing techniques can be utilized to semantically enhance the information retrieval component. Further lexical-semantic knowledge is extracted from resources such as WordNet, Wikipedia [8], and Wiktionary [9]. In the present paper, we have not yet used any semantic information retrieval (IR) techniques in the answer retrieval. Instead, we focus on applying a state of the art IR system to real life user questions from social Q&A sites. We find out that adding elaborate question focus detection techniques to the question processing module is an essential pre-requisite for effective answer retrieval from social media content. Finally, automatic quality assessment is a fundamental technology that has to be applied to open-content QA in educational contexts.

3. Experiments in Question Subjectivity Classification

Most open-domain QA systems apply question and expected answer type classification for adapting different strategies to different types of questions [10]. Despite the large body of work focussing on factoid question type classification, relatively little research

has been done for analysing opinion seeking, i.e., subjective questions [11,12]. These works substantiate that subjective and factoid answers have quite disparate characteristics. Factoid questions typically seek for short informative and objective answers. Subjective questions, on the other hand, seek for longer answers that require distinguishing different opinions in text and presenting similar opinions in an aggregated way.

Some categories of social Q&A sites relevant for the learner's information needs contain a significant amout of subjective questions that require special processing. For instance, consider the following two authentic questions taken from the *homework help* category of the social Q&A site Yahoo!Answers[3] (YA):

- *Should religion be discussed in public schools? I'm doing a research for my school about this topic and i really don't know much about it. I have to present this in front of the WHOLE class but i feel like an ignorant...can you help please?*
- *What's an example of ignorance in our society today? i've gotta write a paper on this and i need more examples!*

In both questions, the learners seek answers containing different perspectives. Answers to these questions can vary based on personal opinions, judgments, and experiences. Unlike factoid answers, we cannot say that one answer is superior to another. Therefore, instead of a single best answer, learners should be presented with an overview of different perspectives.

For question subjectivity classification experiments we compiled a dataset from YA questions and answers[4] from 4 different categories, i.e. *Teaching (100 questions) , Homework Help (101 questions), Books&Authors (101 questions)* and *Environment (68 questions)*. We employ two human annotators and compute the Kappa statistics for inter-annotator agreement on subsets of the data. The annotators were asked to annotate each question as either seeking for opinions or as seeking for factual information. On 134 questions from the *Teaching (42 questions), Books&Authors (46 questions),* and *Environment (46 questions)* categories, the annotators reach a Kappa of 0.78 indicating sufficient agreement. Therefore, the rest of the data was annotated by one annotator only. The distribution of the subjective questions for 4 categories is as follows (in percentage): Teaching (92%), Homework Help (47%), Books&Authors (94%), and Environment (42%).

We propose an unsupervised lexicon-based approach for question subjectivity classification. We split the data into two subsets maintaining the same proportions from each category: 176 questions for training and 189 questions for testing. The approach utilizes two knowledge sources, hereafter *subjectivity clues*, that were manually crafted based on the analysis of training data: (i) a lexicon with 137 single and 69 multi-word entries, e.g. *what do you think, your favorite, better than*, and (ii) a list of 14 part-of-speech (POS) sequences, e.g. *adj conj adj, art v pr adv*. For each clue instance we compute a subjectivity score ss_c as $ss_c = \sum_{i=1}^{k} 2^i$ where k is the number of unigrams in a clue. Then we calculate a subjectivity score ss_q for each question as $ss_q = \sum_{i=1}^{j} ss\_c_i$ where j is the number of subjectivity clues in a question. As this approach may boost the subjectivity score for longer questions, we emprically set thresholds based on the number of sentences in a question. Questions with less than 4 sentences[5] are classified as subjective

[3] http://answers.yahoo.com/

[4] http://ir.mathcs.emory.edu/shared/

[5] Questions on the YA platform often contain detailed descriptions of the problem at hand, as shown in the examples given above.

if $ss_q \geq 3$, and questions with more than 4 sentences are classified as subjective if

$$ss_q \geq \lfloor 5 - \frac{1}{2}n \rfloor$$

where n is the number of sentences in a question. Using this approach, we achieve an F-measure of 0.86 over 189 test questions, and an F-measure of 0.88 over the whole set of 365 questions. In the work by [13], a supervised machine learning approach SVM with linear kernel is utilized to predict question subjectivity on YA data based on a set of characters and mixed word and POS n-gram features. They conduct experiments using the text of the question, the text of the best answer, the text of all answers, the text of both the question and the best answer, and the text of the question with all answers. They achieve 0.742 macro-averaged F-measure based on the combination of the question text and all answers when these are treated as separate feature spaces, and 0.72 based on the question text only. As we only have question texts in an online situation, our approach is based on the question's text only and does not depend on the answers.

Besides question analysis, answer retrieval is the other major component in our educational QA system. In the following section, we describe our current approach to answer retrieval which relies on information retrieval.

4. Question Answering as Information Retrieval

Educational QA has to deal with a huge variety of heterogeneous information sources, such as Wikipedia, blogs, slides of scientific presentations, or social Q&A sites. The search for exact answers in long documents, such as Wikipedia articles, requires a sophisticated answer extraction component. However, answer extraction is known as one of the fundamental problems in QA due to the vocabulary gap between questions and answers [14,15]. At the same time, social Q&A sites contain large repositories of previously asked questions and their corresponding human-generated answers, which do not necessarily require any answer extraction from scratch. This way, we can explore information retrieval methods operating on existing Q&A repositories as an alternative solution to QA [16,17,18].

In our initial experiments, we focussed on assessing the performance of the Lucene text search library [19] in our educational QA system. The questions consist of 25 real user questions randomly selected from the social Q&A site Answerbag[6], e.g. *When should one use COMP FIELDS in COBOL , and what is their use*.

The document collection employed in information retrieval consists of Question-Answer pairs extracted from Yahoo!Answers computer related categories. We follow the approach described in [17] and index separate fields in the document collection - category, question and answer field. The final relevance score is computed as the weighted sum of relevance scores after retrieval on each of the fields[7]. Following [17], one of the authors manually classified each retrieved document in one of the three categories *answer* (the document contains an exact answer to the original question), *interesting* (the

[6] http://www.answerbag.com/

[7] The corresponding weights: question 0.5, category 0.3, answer 0.2

document does not contain the exact answer, but contributed relevant information necessary to answer the question), and *irrelevant*. To measure the system performance, we apply two metrics: Success@n ($S@n$) [17] and Mean Reciprocal Rank ($MRR@n$) [20]. Success@n is defined as the number of questions with at least one correct answer in the top n results. The reciprocal rank (RR) is the inverse of the rank of the highest ranking answer, while the MRR measure is the mean RR across all queries.

The experimental results for three different n values are presented in Table 1 (without parentheses). No performance increase can be observed for $n = 20$ as compared to $n = 10$, while the performance increase in $n = 10$ as compared to $n = 5$ is more prominent for the $S@n$ measure. The $S@n$ and $MRR@n$ measures are stable for all values of n in the "Answer & Interesting" category which shows that if no interesting answer occurs in the top 5 answers retrieved for a given question, then only irrelevant answers are retrieved. An analysis of IR results revealed that the system performance deteriorates due to the missing linguistic preprocessing of questions, which are often quite long and unclear. For example, for the question *When I double click on IE it doesn't open anymore, but when I go to winamp for example and click get more skins, it opens a window right away, how can I fix this? Even opera started doing it.*, the focus of the search should be *When I double click on IE it doesn't open anymore*, while Lucene converts it to the query *click click open open double ie anymore go winamp example get skin window right away can fix even opera start* and gives the highest weighting to *double click*.

Table 1. QA as information retrieval results based on 25 Answerbag questions. Figures in parentheses correspond to manually focussed questions.

	S@5	**S@10**	**S@20**	**MRR@5**	**MRR@10**	**MRR@20**
Answer	0.24 (0.36)	0.32 (0.4)	0.32 (0.4)	0.22 (0.28)	0.23 (0.29)	0.23 (0.29)
Answer & Interesting	0.92 (0.92)	0.92 (0.92)	0.92 (0.92)	0.66 (0.66)	0.66 (0.66)	0.66 (0.66)

We conclude that more sophisticated question analysis addressing question focus detection is required to improve answer retrieval. Therefore, we performed an additional experiment, whereby the real questions were manually converted to more focussed questions as in the example given above. The results of answer retrieval for manually focussed questions are given in parentheses in Table 1. Manually focussed questions improve the results for the "Answer" category. In future work, we will therefore explore methods as proposed in [21] to perform question preprocessing.

5. Conclusions

In this paper, we thoroughly analysed the requirements for an educational QA system operating on social media content. We showed that such a system requires the employment of advanced NLP technologies in addition to standard QA components. The resulting system architecture should be designed to deal with heterogeneous and error-prone data. It should include a sophisticated question analysis component capable of question subjectivity classification (Section 3), as the number of subjective questions in social media is high. Furthermore, it should include capabilities for generating high-quality questions from low-quality questions [6]. Finally, a lot of textually encoded information in social

Q&A repositories can be re-used by utilizing question paraphrase identification [1] and advanced information search technologies (Section 4).

We showed that the subjectivity of questions for a set of categories in the domain of educational QA can be reliably annotated by human coders and proposed a simple lexicon-based approach to identification of subjective questions yielding promising results. We found that question preprocessing, especially question type and question focus analysis, are vital to the success of QA systems operating on social media content.

Finally, future work in educational QA will have to extensively address automatic quality assessment which becomes crucial especially in the learning domain. Previous studies [22,23] showed that the quality of textual documents can be reliably measured using machine learning and natural languages processing techniques. Adapting these techniques to different discourse types in educational QA will require further research. Another area that requires research attention is answer processing. Dealing with dozens and hundreds of answers to individual questions in social Q&A sites calls for multi-document summarization techniques tuned to serve the needs of educational QA. Furthermore, a system operating on several types of discourse will have to find optimal ways of presenting the answers derived from distinct information sources to the learner.

Acknowledgements

This work has been supported by the Emmy Noether Program of the German Research Foundation (DFG) under the grant No. GU 798/3-1, and by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under the grant No. I/82806.

References

[1] D. Bernhard and I. Gurevych, "Answering Learners' Questions by Retrieving Question Paraphrases from Social Q&A Sites," in *Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications in conjunction with ACL 2008*, (Columbus, Ohio, USA), pp. 44–52, June 19 2008.

[2] E. Hovy, L. Gerber, U. Hermjakob, M. Junk, and C. Lin, "Question Answering in Webclopedia," in *Proceddings of TREC'2000*, 2000.

[3] S. Ravi, J. Kim, and E. Shaw, "Mining On-line Discussions: Assessing Technical Quality for Student Scaffolding and Classifying Messages for Participation Profiling," in *Educational Data Mining, Supplementary Proceedings of the 13th International Conference of Artificial Intelligence in Education* (C. Heiner, N. Heffernan, and T. Barnes, eds.), (Marina del Rey, CA. USA.), July 2007.

[4] D. Feng, E. Shaw, J. Kim, and E. Hovy, "An Intelligent Discussion-Bot for Answering Student Queries in Threaded Discussions," in *Proceedings of the 11th international conference on Intelligent user interfaces (IUI'06)*, pp. 171–177, 2006.

[5] V. Rus, Z. Cai, and A. C. Graesser, "Experiments on Generating Questions About Facts," in *Proceedings of CICLing* (A. F. Gelbukh, ed.), vol. 4394 of *Lecture Notes in Computer Science*, pp. 444–455, Springer, 2007.

[6] K. Ignatova, D. Bernhard, and I. Gurevych, "Generating High Quality Questions from Low Quality Questions," in *Proceedings of the Workshop on the Question Generation Shared Task and Evaluation Challenge*, (NSF, Arlington, VA), Sep 2008. `http://www.cs.memphis.edu/~vrus/questiongeneration//8-Ignatova-QG08.pdf`.

[7] K. Ignatova, C. Toprak, D. Bernhard, and I. Gurevych, "Annotating Question Types in Social Q&A Sites," in *GSCL Symposium "Language Technology and eHumanities"*, 2009.

[8] T. Zesch, I. Gurevych, and M. Mühlhäuser, "Analyzing and Accessing Wikipedia as a Lexical Semantic Resource," in *Data Structures for Linguistic Resources and Applications* (G. Rehm, A. Witt, and L. Lemnitzer, eds.), pp. 197–205, Tuebingen, Germany: Gunter Narr, Tübingen, 2007.

[9] T. Zesch, C. Mller, and I. Gurevych, "Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary," in *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, (Marrakech, Morocco), May 2008.

[10] J. Prager, J. Chu-Carroll, and K. Czuba, "Statistical answer-type identification in open-domain question answering," in *Proceedings of the second international conference on Human Language Technology Research*, (San Francisco, CA, USA), pp. 150–156, Morgan Kaufmann Publishers Inc., 2002.

[11] V. Stoyanov, C. Cardie, and J. Wiebe, "Multi-Perspective Question Answering Using the OpQA Corpus," in *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HTL-EMNLP 2005)*, pp. 923–930, 2005.

[12] B. Li, Y. Liu, A. Ram, E. V. Garcia, and E. Agichtein, "Exploring question subjectivity prediction in community QA," in *SIGIR '08: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (New York, NY, USA), pp. 735–736, ACM, 2008.

[13] B. Li, Y. Liu, A. Ram, E. V. Garcia, and E. Agichtein, "Exploring question subjectivity prediction in community QA," in *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, (New York, NY, USA), pp. 735–736, ACM, 2008.

[14] A. Berger, R. Caruana, D. Cohn, D. Freitag, and V. Mittal, "Bridging the lexical chasm: statistical approaches to answer-finding," in *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, 2000.

[15] S. Riezler, A. Vasserman, I. Tsochantaridis, V. Mittal, and Y. Liu, "Statistical Machine Translation for Query Expansion in Answer Retrieval," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, (Prague, Czech Republic), pp. 464–471, Association for Computational Linguistics, June 2007.

[16] J. Jeon, B. W. Croft, and J. H. Lee, "Finding similar questions in large question and answer archives," in *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, (New York, NY, USA), pp. 84–90, ACM, 2005.

[17] V. Jijkoun and M. de Rijke, "Retrieving answers from frequently asked questions pages on the web," in *CIKM '05*, (New York, NY, USA), pp. 76–83, ACM, 2005.

[18] X. Xue, J. Jeon, and W. B. Croft, "Retrieval models for question and answer archives," in *SIGIR*, pp. 475–482, 2008.

[19] O. Gospodnetic and E. Hatcher, *Lucene in Action*. Manning Publications Co., 2005.

[20] E. Voorhees, "The TREC-8 question answering track report," in *Proceedings of the 8th Text Retrieval Conference*, (Gaithersburg, Maryland, USA), pp. 77–82, 1999.

[21] H. Duan, Y. Cao, C.-Y. Lin, and Y. Lu, "Searching Questions by Identifying Question Topic and Question Focus," in *Proceedings of ACL 2008*, (Columbus, Ohio, USA), June 2008.

[22] M. Weimer, I. Gurevych, and M. Mühlhäuser, "Automatically Assessing the Post Quality in Online Discussions on Software," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, (Prague, Czech Republic), pp. 125–128, Association for Computational Linguistics, June 2007.

[23] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, "Finding high-quality content in social media," in *WSDM '08: Proceedings of the international conference on Web search and web data mining*, (New York, NY, USA), pp. 183–194, ACM, 2008.