Understanding and Summarizing Answers in Community-Based Question Answering Services

Yuanjie Liu[1], Shasha Li[2], Yunbo Cao[1,3], Chin-Yew Lin[3], Dingyi Han[1], Yong Yu[1]

[1]Shanghai Jiao Tong University, Shanghai, China, 200240
{lyjgeorge,handy,yyu}@apex.sjtu.edu.cn

[2]National University of Defense Technology, Changsha, China, 410074
Shashali@nudt.edu.cn

[3]Microsoft Research Asia, Beijing, China, 100080
{yunbo.cao,cyl}@microsoft.com

Abstract

Community-based question answering (cQA) services have accumulated millions of questions and their answers over time. In the process of accumulation, cQA services assume that questions always have unique *best answers*. However, with an in-depth analysis of questions and answers on cQA services, we find that the assumption cannot be true. According to the analysis, at least 78% of the cQA best answers are reusable when similar questions are asked again, but no more than 48% of them are indeed the unique *best answers*. We conduct the analysis by *proposing taxonomies for cQA questions and answers*. To better reuse the cQA content, we also *propose applying automatic summarization techniques to summarize answers*. Our results show that question-type oriented summarization techniques can improve cQA answer quality significantly.

1 Introduction

Community-based question and answering (cQA) service is becoming a popular type of search related activity. Major search engines around the world have rolled out their own versions of cQA service. Yahoo! Answers, Baidu Zhidao, and Naver Ji-Sik-In[1] are some examples.

In general, a cQA service has the following workflow. First, a question is posted by the asker in a cQA service and then people in the community can answer the question. After enough number of answers are collected, a best answer can be chosen by the asker or voted by the community. The resulting question and answer archives are large knowledge repositories and can be used to complement online search. For example, Naver's Ji-Sik-In (Knowledge iN) has accumulated about 70 million entries[2].

In an ideal scenario, a search engine can serve similar questions or use best answers as search result snippets when similar queries are submitted. To support such applications, we have to assume the best answers from cQA services are good and relevant answers for their pairing questions. However, the assumption might not be true as exemplified by the following examples.

Question Title	Which actress has the most seductive voice?..could range from a giggly goldie hawn..to a sultry anne bancroft?
Question Description	or any other type of voice that you find alluring. ..
Best Answer (Polls & Surveys)	Fenella Fielding, wow!!!!
Best Answer (Movies)	i think joanna lumlley has a really sexy voice

Table 1. Same Question / Different Best Answers

Question Title	Does anyone know of any birthdays coming up soon?
Question Description	Celerities, people you know, you? Anyway I need the name and the date. If you want to know it is for my site, http://www.jessicaparke2.piczo.com... and that is not site advertising.
Answer	Novembers Are: Paul Dickov nov 1st Nelly (not furtado) nov 2nd …
Best Answer	Check imdb.com, they have this celebrity birthdays listed.

Table 2. Question with Alternative Answers

Table 1 presents a question asking community opinions about "*who is the actress has the most seductive voice*". The asker posted the same question twice at different Yahoo! Answers categories: one in *Polls & Surveys* and one in *Mov-*

[1] Yahoo! Answers: answers.yahoo.com; Baidu Zhidao: zhidao.baidu.com; Naver Ji-Sik-In: kin.naver.com

[2] www.iht.com/articles/2007/07/04/technology/naver.php

ies. Two different best answers were chosen by the same asker due to non-overlapping of answers. Table 2 shows another example, it asks about "*the coming birthdays of stars*". The best answer chosen by the asker is very good because it provides useful URL information where the asker can find her answers. However, other answers listed a variety of birthdays of stars that also answered the question. These two examples indicate that the conventional cQA policy of allowing askers or voters to choose best answers might be working fine with the purpose of cQA but it might not be a good one if we want to reuse these *best answers* without any post-processing.

To find out what might be the alternatives to the *best answer*s, we first carried out an in-depth analysis of cQA data by *developing taxonomies for questions and answers*. Then we propose summarizing answers in a consideration of *question type*, as the alternative to the *best answers.* For example, for the '*actress voice*' question, a summary of different people's opinions ranked by popularity might be a better way for expressing the question's answers. Similar to the '*actress voice*' question, the '*celebrity birthday*' question does not have a fix set of answers but is different from the '*actress voice*' question that its answers are facts not opinions. For fact-based open ended questions, combining different answers will be useful for reuse of those answers.

The rest of this paper is arranged as follows. We review related work in Section 2. We develop a framework for answer type taxonomy in Section 3 and a cQA question taxonomy in Section 4. Section 5 presents methods to summarize cQA answers. Finally, we conclude this paper and discuss future work in Section 6.

2 Related Work

Previous research on cQA (community-based Question and Answering) domain focused on three major areas: (1) how to find similar questions given a new question (Jeon et al. 2005a; Jeon et al., 2005b), (2) how to find experts given a community network(Liu et al., 2005; Jurczyk & Agichtein, 2007), and (3) how to measure answer quality and its effect on question retrieval. The third area of focus is the most relevant to our research. Jeon et al. (2006)'s work on assessing cQA answer quality is one typical example. They found that about 1/3 of the answers among the 1,700 Q&A pairs from Naver.com cQA data have quality problems and approximately 1/10 of them have bad answers[3]. They used 13 non-textual features and trained a maximum entropy model to predict answer quality. They showed that retrieval relevance was significantly improved when answer quality measure was integrated in a log likelihood retrieval model.

As mentioned in Section 1, cQA services provide an alternative way for users to find information online. Questions posted on cQA sites should reflect users' needs as queries submitted to search engines do. Broder (2002) proposed that search queries can be classified into three categories, i.e. *navigational*, *informational*, and *transactional*. Ross and Levinson (2004) suggested a more elaborated taxonomy with five more subcategories for informational queries and four more subcategories for resource (transactional) queries. In open-domain question answering research that automatic systems are required to extract exact answers from a text database given a set of factoid questions (Voorhees and M. Ellen, 2003), all top performing systems had incorporated question taxonomies (Hovy et al., 2001; Moldovan et al., 2000; Lytinen et al., 2002; Jijkoun et al., 2005). Based on the past experiences from the annual NIST TREC Question and Answering Track[4] (TREC QA Track), an international forum dedicating to evaluate and compare different open-domain question answering systems, we conjecture that a cQA question taxonomy would help us determine what type of best answer is expected given a question type.

Automatic summarization of cQA answers is one of the main focuses of this paper. We propose that summarization techniques (Hovy and Lin, 1999; Lin and Hovy, 2002) can be used to create cQA answer summaries for different question types. Creating an answer summary given a question and its answers can be seen as a multi-document summarization task. We simply replace documents with answers and apply these techniques to generate the answer summary. The task has been one of the main tasks the Document Understanding Conference[5] since 2004.

3 A Framework for Answer Type

To study how to exploit the best answers of cQA, we need to first analyze cQA answers. We would like to know whether the existing best answer of a specific question is good for reuse. If not, we

[3] Answers in Jeon el al.'s work were rated in three levels: good, medium, and bad.
[4] http://trec.nist.gov/data/qamain.html
[5] http://duc.nist.gov

want to understand why and what the alternatives are. We will refer to the 'best answers' selected by cQA askers or voters as BA henceforth to differentiate it with best answers annotated or automatically generated in our experiments.

We made use of questions from Yahoo! Answers for developing and testing our framework for answer type. There are over 1,000 hierarchical categories in Yahoo! Answers. By manually examining 400 randomly selected questions from the 4 most popular top Yahoo! Answers categories (100 questions from each category) – *Entertainment & Music (E&M)*, *Society & Culture (S&C)*, *Health*, and *Computers & Internet (C&I)*, we developed a cQA answer type taxonomy based on *the principle of BA reusability* that determines a BA's answer type based on "*if the BA can be reused or not when a question that is similar to the BA's question is asked again*".

One of the authors carried out this manual exercise and developed the initial answer taxonomy. The taxonomy was then modified through discussions among the authors. We asked three annotators to do the annotation. We assigned the category label that was agreed by at least two annotators. If none of the three annotators agreed on a single category label, one of the authors made the final decision. The answer type taxonomy is described in this section and we discuss the question type and the relation with answer taxonomy in next section.

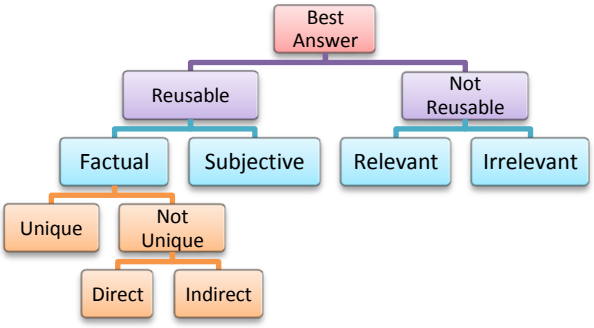

Figure 1. cQA Services BA Type Taxonomy.

Figure 1 shows the resulting answer type taxonomy. It first divides BA into two categories: *Reusable* and *Not Reusable*. A *Reusable* BA means that it can be reused as the best answer if a similar question to its question is asked again; while a *Not Reusable* BA means it cannot be reused. The *Reusable* BA is further divided into *Factual* and *Subjective*. A *Factual* BA is a fact that can be used as the best answer; while a *Subjective* BA is one of the opinions that can be used as the best answer.

The *Factual* BA type has two subtypes: *Unique* and *Not Unique*. A *Unique* BA has only a unique best answer to its question and no other answer add more information; while a *Not Unique* BA has other alternative best answers. The *Not Unique* BA type is divided into two subtypes: *Direct* and *Indirect*. A *Direct* BA answers its question directly; while an *Indirect* BA answers its question through inference. For example, the question mentioned in section 1 has the *Indirect* BA which gives a website reference, while there is also a *Direct* answer just gives the birthday lists.

A *Subjective* BA answers questions that look for opinions or recommendations. For example, a question asked "Which is the best sci-fi movie?" Each answerer would have his own idea.

The *Not Reusable* BA has two subtypes: *Relevant* and *Irrelevant*. A *Relevant* BA could not be used as a best answer to its question but it is relevant to its question, for example, a question asked "*Why was "I Can't Hate You Anymore" by Nick Lachey so shortlived?*" A *Relevant* BA said "*I'm not sure where you live, but in NJ, especially South Jersey, that song was played out…*", this answer is relevant but without knowing the asker's location which does not really answer the question; an *Irrelevant* BA could not be used as a best answer to its question and it is irrelevant to its question. The BA "*It appears that the question period has expired. If an answer has been given that meets your needs, please pick a 'best answer'.*" of the question "*how to forward an email without showing the email addresses in the To box*" is in this case.

Answer Type	**C&I**	**E&M**	**Health**	**S&C**
Unique	47%	28%	48%	13%
Direct	28%	7%	30%	18%
Indirect	9%	3%	5%	2%
Factual Total	**84%**	**38%**	**83%**	**33%**
Subjective	4%	40%	7%	50%
Reusable Total	**88%**	**78%**	**90%**	**83%**
Relevant	3%	1%	1%	0%
Irrelevant	9%	21%	9%	17%
Not Reusable Total	**12%**	**22%**	**10%**	**17%**

Table 3. Distribution of Answer Type

Table 3 shows the distribution of Answer types on four categories. Unique answers are no more than 48%. The C&I and the Health categories tend to have more factual BAs than other two categories.

Among the four categories, S&C answers are mostly not unique and have a high percentage (50%) of subjective answers. This indicates that the one BA per cQA question chosen by its asker or voters is not good enough for reuse as the best answer. However, we might be able to apply au-

tomatic summarization techniques to create summarized answers for at least half of reusable (but not unique) answers. We provide some possible solutions in Section 5.

Category	Percentage
Computer & Internet	18%
Entertainment & Music	17%
Health	21%
Society & Culture	20%

Table 4. Disagreement on Answer Type

Table 4 shows the percentage of questions over which none of the three annotators agreed on a single category label. The results show that the question taxonomy developed above is pretty stable (over at least 79% questions).

4 A CQA Question Taxonomy

As we were developing our answer type taxonomy, we often could not solely rely on answers themselves and had to consider their questions as well. As we discussed in Section 2, the type of question would help us determine the expected best answer types.

Rose and Levinson's (2004) taxonomy of search engine queries has similar goal to ours though their taxonomy was developed to classify search engine queries. Instead of starting from scratch, we followed the basic hierarchy of R&L's taxonomy and made some modifications to accommodate the particular of cQA services.

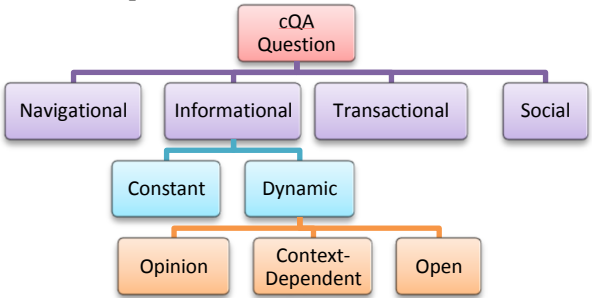

Figure 2. Question Type Taxonomy

Figure 2 shows the resulting question taxonomy. We retain Broder's taxonomy at top levels and propose a new *Social* category. *Navigational*, *Informational* and *Transactional* are defined similar as in Broder's taxonomy while *Social* category consists of questions that do not intend to get an answer but just were used to elicit interaction with people in cQA services.

Navigational category contains questions seeking URLs of specific websites that the asker would like to visit, for example, "*Does anybody know the fan sites of Hannah Montana?*"

Transactional category contains questions tend to get resources. A typical one is "*Is there a computer program that lets you create a planet?*"

For *Information* category, we first divide it into two subcategories: *Constant* and *Dynamic*. *Constant* questions have a fixed or stable set of answers while *dynamic* questions do not. This dichotomy of informational category is to support our intention to establish intuitive mapping between the question taxonomy and the answer taxonomy. *Constant* question type is similar to R&L's closed query type. An example constant question is "*Which country has the largest population?*" but "*What is the population of China?*" would be a dynamic question.

For *Dynamic* category, we define three subcategories: *Opinion*, *Context-Dependent* and *Open*. *Opinion* questions are those asking for opinions. Questions in this category seek opinions from people in cQA communities about what they think of some people, some events, or some objects. "*Is Microsoft Vista worth it?*" is an example. *Context-dependent* questions are those questions having different answers according to the different context. For example, the question "*What is the population of China?*" should have different answers according to the different date. *Open* category contains questions asking for some facts or methods. The questions usually have a variety of answers or their answer themselves may have unconstrained depth. The question "*Can you list some birthdays of stars in the coming week?"* is an example. This essentially follows R&L's open query category. It also includes what is not covered by the opinion and context-dependent categories.

The new *Social* category is specific to cQA services. Questions in this category do not intend to get an answer. These questions include telling jokes and expressing askers' own ideas. Essentially, askers treat cQA service as chatting rooms or online forums. The question "*Why do so many lazy people come on here simply just to ask...?*" together with the question description "*how to become a hacker? It really isn't that hard to do a google search…hopefully some of the people that will continue to ask, will click the link below so they can give up faster…* " actually is expressing a negative sentiment towards a number of people who asked how to become a hacker.

Table 5 shows the distribution of different question types on 4 different Yahoo! Answers categories. We observe that constant questions only occupy 11% ~ 20% while navigational questions are even fewer such that they do not occur in the sample questions. This is reasonable since people very likely would be able to use search engines to discover answers of naviga-

tional and constant questions. They do not have to ask these types of question on community-based question answering services. On the contrary, open and opinion questions are frequently asked, it ranges from 56%~83%.

Question Type	C&I	E&M	Health	S&C
Navigational Total	**0%**	**0%**	**0%**	**0%**
Constant	15%	20%	15%	11%
Opinion	8%	37%	16%	60%
Context Dependent	0%	1%	1%	0%
Open	59%	19%	67%	18%
Dynamic Total	**67%**	**57%**	**84%**	**78%**
Informational Total	**82%**	**77%**	**99%**	**89%**
Transactional Total	**14%**	**8%**	**0%**	**1%**
Social Total	**4%**	**15%**	**1%**	**10%**

Table 5 Distribution of Question Type

Intersection Number	UNI	DIR	IND	SUB	REL	IRR
Navigational	0	0	0	0	0	0
Constant	**48**	9	3	0	1	0
Open	**51**	**62**	13	15	5	17
Context-dep	0	0	1	0	0	1
Opinion	15	13	1	**84**	0	8
Transactional	**10**	**7**	4	1	0	1
Social	0	0	0	1	0	**29**

Table 6. Question Answer Correlation

Table 6 (UNI: *unique*, DIR: *direct*, IND: *indirect*, SUB: *subjective*, REL: *relevant*, IRR: *irrelevant*) gives the correlation statistics of question type vs. answer type. There exists a strong correlation between question type and answer type. Every question type tends to be associated with only one or two answer types (bold numbers in Table 6).

5 Question-Type Oriented Answer Summarization

Since the BAs for at least half of questions do not cover all useful information of other answers, it is better to adopt post-processing techniques such as answer summarization for better reuse of the BAs. As observed in the previous sections, answer types can be basically predicted by question type. Thus, in this section, we propose to use multi-document summarization (MDS) techniques for summarizing answers according to question type. Here we assume that question type can be determined automatically. In the following sub-sections, we will focus on the summarization of answers to *open* or *opinion* questions as they occupy more than half of the cQA questions.

5.1 Open Questions

Algorithm: For *open* questions, we follow typical MDS procedure: topic identification, interpretation & fusion, and then summary generation (Hovy and Lin, 1999; Lin and Hovy, 2002). Table 7 describes the algorithm.

1. Employ the clustering algorithm on answers
2. Extract the noun phrases in each cluster, using a shallow parser.[6]
3. For each cluster and each label (or noun phrase), calculate the score by using the Relevance Scoring Function:

$$\sum_{w} p(w|\theta)PMI(w,l|C) - D(\theta|C)$$

Where θ is the cluster, w is the word, l is the label or noun phrase, C is the background context which is composed of 5,000 questions in the same category, p(·) is conditional probability, PMI(·) is pointwise mutual information, and D(·) is KL-divergence

4. Extract the key answer which contains the noun phrase that has the highest score in each cluster
5. Rank these key answers by cluster size and present the results.

Table 7. Summarization Algorithm(Open-Type)

In the first step, we use a bottom-up approach for clustering answers to do topic identification. Initially, each answer forms a cluster. Then we combine the most similar two clusters as a new cluster if their similarity is higher than a threshold. This process is repeated until no new clusters can be formed. For computing similarities, we regard the highest cosine similarity of two sentences from two different clusters as the similarity of the two clusters. Then we extract salient noun phrases, i.e. cluster labels, from each cluster using the first-order relevance scoring function proposed by Mei et al. (2007), (step 2,3 in Table 7). In the fusion phase (step 4), these phrases are then used to rank answers within their cluster. Finally in the generation phase (step 5), we present the summarized answer by extracting the most important answer in every cluster and sort them according to the cluster size where they come from.

Case Example: Table 8 presents an example of summarization results of open-type questions. The question asks how to change Windows XP desktop to Mac style. There are many softwares providing such functionalities. The BA only lists one choice – the StarDock products, while other answers suggest Flyakite and LiteStep. The automatic summarized answer (ASA) contains a variety of for turning Windows XP desktop into Mac style with their names highlighted as cluster labels. Compared with manually-summarized answer (MSA), ASA contains most information of MSA while retains similar length with BA and MSA.

5.2 Opinion Questions

Algorithm: For *opinion* questions, a comprehensive investigation of this topic would be beyond the scope of this paper since this is still a field

[6] http://opennlp.sourceforge.net

Question (http://answers.yahoo.com/question/?qid=1005120801427)
What is the best way to make XP look like Mac osX?
Best Answer Chosen
I found the best way to do this is to use WindowsBlinds. A program that, if you use the total StarDock, package will allow you to add the ObjectBar in addition to changed the toolbars to be OS X stylized. If you want added functionality you can download programs off the internet that will mimic the Expose feature which will show you a tiled set of all open windows. Programs that will do this include: WinPlosion, Windows Exposer, and Top Desk
Auto-summarized Answer
LiteStep: An additional option is LiteStep - a "Shell Replacement" for Windows that has a variety of themes you can install. Undoubtedly there are various Mac OSX themes avaialable for LiteStep. I have included a source to a max osx theme for Litestep at customize.org. **Flyakite:**Flyakite is a transformation pack and the most comprehensive in terms of converting an XP system's look to that of an OS X system, google it up and you should find it, v3 seems to be in development and should be out soon. **Window Blinds:**http://www.stardock.com/products/windowb...
Manually-summarized Answer
One way is to use WindowsBlinds. The package will allow you to add the ObjectBar for changing to the OSX theme. You can also make added functionality of Expose feature by downloading the programs like WinPlosion, Windows Exposer and Top Desk. The URL of it is http://www.stardock.com/products/windowblinds/. Another option is to use Flyakite which is a transformation pack. The third Option is the LiteStep, it is a "Shell Replacement" for windows that has a variety of Mac OSX tehmes you can install. The url is http://litestep.net and I have included a source of Mac OS theme for Litestep at http://www.customize.org/details/33409.

Table 8. Summary of Open-Question

under active development (Wiebe et al., 2003; Kim and Hovy, 2004). We build a simple yet novel opinion-focused answer summarizer which provides a global view of all answers. We divide opinion questions into two subcategories. One is *sentiment*-oriented question that asks the sentiment about something, for example, "*what do you think of* ...". The other is *list*-oriented question that intends to get a list of answers and see what item is the most popular.

For sentiment-oriented questions, askers care about how many people support or against something. We use an opinion word dictionary[7], a cue phrase list, a simple voting strategy, and some heuristic rules to classify the sentences into *Support*, *Neutral*, or *Against* category and use the overall attitude with key sentences to build summarization. For list-oriented questions, a simple counting algorithm that tallies different answers of questions together with their supporting votes would be good answer summaries. Details of the algorithm are shown in Table 9, 10.

Case Example: Table 11 presents the summarization result of an sentiment-oriented question, it asks "whether it is strange for a 16-year child to talk to a teddy bear?", the BA is a negative response. However, if we consider all answers, we find that half of the answers agree but another half of them disagree. The distribution of different sentiments is similar as MSA. Table 12 shows the summarization result of a list-oriented question, the question asks "*what is the best sci-fi movie?*" The BA just gives one choice "Independence day" while the summarized answer gives a list of best sci-fi movies with the number of supporting vote. Though it is not complete compared with MSA, it contains most of the options which has highest votes among all answers.

1. Employ the same cluster procedure of Open-Type question.
2. If an answer begins with negative cue phrase (e.g. "*No, it isn't*" etc.), it is annotated as *Against*. If a response begins with positive cue phrase (e.g. "*Yes, it is*" etc.), it is annotated as *Support*.
3. For a clause, if number of positive sentiment word is larger than negative sentiment word, the sentiment of the clause is *Positive*. Otherwise, the sentiment of the clause is *Negative*.
4. If there are negative indicators such as "*don't/never/...*" in front of the clause, the sentiment should be reversed.
5. If number of negative clauses is larger than number of positive clauses, the sentiment of the answer is *Negative*. Otherwise, the sentiment of the answer is *Positive*.
6. Denote the sentiment value of question as s(q), the sentiment value of an answer as s(a), and then the final sentiment of the answer is logical AND of s(q) and s(a)
7. Present key sentiments with attitude label

Table 9. Summarization Algorithm (Sentiment-Opinion)

1. Segment the answers into sentences
2. Cluster sentences by using similar process in open-type
3. For each cluster, choose the key sentence based on mutual information between itself and other sentences within the cluster
4. Rank the key sentences by the cluster size and present them ogether with votes

Table 10. Summarization Algorithm (List-Opinion)

Question (http://answers.yahoo.com/question/?qid=1006050125145)
I am 16 and i stil talk to my erm..teddy bear..am i wierd???
Best Answer Chosen
not at all i'm 14 and i too do that
Auto-summarized Answer
Support **A:** It's might be a little uncommon for a 16 year old to talk to a teddy bear but there would be a serious problem if you told me that your teddy bear answered back as you talked to him!!:) **A:** I slept with my teddy bear until I graduated. Can't say that I ever had a conversation with him, but if I had I'm sure he would've been a very good listener. **Against** **A:** i talk to a seed im growing .. its not weird .... :) **A:** No, you're not weird.....you're Pratheek! :D **A:** no, i like to hold on to my old memories too. i do it sometimes too. **A:** It will get weird when he starts to answer back! **A:** not really. it depends how you talk i mean not if you talk to it like its a little kid like my brother does. Overall Attitude: **Support 5 / Neutral 1 / Against 5**
Manually-summarized Answer
support (vote 4) **neutral** (vote 2) **against** (vote 5) reasons: i like to hold on to my old memories too. (vote 1) I slept with my teddy bear until I graduated. (vote 1) i'm 14 and i too do that (vote 1)

Table 11. Summary of Sentiment-Opinion Question

[7] Inquirer dictionary http://www.wjh.harvard.edu/~inquirer.

Question (http://answers.yahoo.com/question/?qid=20060718083151AACYQJn)
What is the best sci-fi movie u ever saw?
Best Answer Chosen
Independance Day
Auto-summarized Answer
star wars (5) Blade Runner (3) fi movie has to be Night of the Lepus (2) But the best "B" sci (2) I liked Stargate it didn't scare me and I thought they did a great job recreating Egypt (3) Independance Day (3)
Manually-summarized Answer
Star Wars (vote 6); The Matrix (vote 3); Independence Day (vote 2); Blade Runner (vote 2); Starship Troopers (vote 2); Alien (vote 2); Alien v.s Predator (vote 1); MST3K (vote 1);

Table 12. Summary of List-Opinion Question

5.3 Experiments

Information Content: To evaluate the effectiveness of automatic summarization, we use the information content criterion for comparing ASA with BA. It focuses on whether ASA or BA contains more useful information to the question. Information point is used in the evaluation. Usually, one kind of solution for *open* questions or one kind of reason for *opinion* questions can contribute one information point. By summing all information points in both ASA and BA, we then can compare which one contains more information. Intuitively, longer texts would contain more information. Thus, when comparing the information content, we limit the length of ASA with several levels to do the evaluation. Take question in Table 8 as an example, the BA just gives one software, which contributes one information point while the ASA lists three kinds of software which contributes three information points. Thus, ASA is considered better than BA.

For each question, we generate 100%, 150%, and 200% BA word-length ASAs. Three annotators are asked to determine whether an ASA is better than, equal to, or worse than its corresponding BA in terms of information content. Voting strategy is used to determine the final label. If three labels are all different, it is labeled as Unknown. We extract 163 *open* questions and 121 *opinion* questions from all four categories by using final question category labels mentioned in Section 4. To make meaningful comparison, questions having unique answers or having only one answer are excluded. After the removal, there are 104 *open* questions and 99 *opinion* questions left for comparison. The results are shown in Table 13.

We are encouraged by the evaluation results that our automatic summarization methods generate better coverage of contents in most of the cases at every answer summary length. We observe a big difference between 100% and 150% answer summaries. It should not be a surprise since a 150% answer summary contains 50% more content than its corresponding BA. While at the 100% length, we still have about 30% ASAs better than BA. Questions which have better ASA than BA usually have a long BA but with little information. Table 14 provides the example. By using summarization, answers that are compact and direct to the question can be included. The results indicate that summary compression technique might be helpful to pack more information in short answers.

Open	**ASA Better**	**BA Better**	**Equal**	**Unknown**
100%	30%	12%	45%	13%
150%	55%	7%	28%	10%
200%	63%	4%	24%	9%
Opinion	**ASA Better**	**BA Better**	**Equal**	**Unknown**
100%	37%	20%	32%	11%
150%	44%	16%	30%	10%
200%	54%	16%	23%	7%

Table 13. Evaluation by Information Content

Q	Why wont japanese characters burn onto the DVD?
BA	man, the answers here are too stupid for hteir own.You are creating a DVD on Western Platform. I take it, you are using an OS that is in English?In order to "view" japanese as part of your filenames, you need your operating system to accept Japanese coding (characters).If you are using Windows, then you will need ot isntall the Japanese character Set for your operating system If you are using MacOS . i have no idea.
100% ASA	**The dvd writer** Probably because your burner, the DVD writer, doesn't support double bytes code, such as Japanese, Korean, and Chinese. Check the supporting language of your software. Or change all the file name in single byte code, like alphabets. man, the answers here are too stupid for hteir own. You are creating a DVD on Western Platform. I take it, you are using an OS that is in English?

Table 14. Examples of 100% ASA

Readability: Besides the information content, we would also like to study the readability of automatic summarized answers. 10 questions (each from *open* and *opinion* category) are extracted and we make both manual summarized answer (MSA) and automatic summarized answer (ASA) for comparison with BA. We used the information content (INFO) and readability (READ) criteria for evaluation. The readability is judged basically by the time for understanding. We make two kinds of comparison: ASA vs. BA and MSA vs. BA. The first one is used to judge whether the current summarization method is better than current cQA scenario. The second one is used as an expectation for how much the summarization methods can be better than BA.

For ASA vs. BA, the results in Table 15 show that all the annotators agree ASAs providing more information content but not being with satisfying readability. For MSA vs. BA, better results in readability can be achieved as Table 16. This suggests that the proposed approach can succeed as more sophisticated summarization techniques are developed.

Open	Annotator 1		Annotator 2		Annotator 3	
ASA	**INFO**	**READ**	**INFO**	**READ**	**INFO**	**READ**
Better	40%	10%	90%	10%	80%	0%
Equal	60%	60%	10%	80%	20%	60%
Worse	0%	30%	0%	10%	0%	40%
Opinion	**Annotator 1**		**Annotator 2**		**Annotator 3**	
ASA	**INFO**	**READ**	**INFO**	**READ**	**INFO**	**READ**
Better	90%	10%	90%	10%	70%	40%
Equal	10%	60%	10%	60%	10%	20%
Worse	0%	30%	0%	30%	20%	40%

Table 15. ASA vs. BA Evaluation

Open	Annotator 1		Annotator 2		Annotator 3	
MSA	**INFO**	**READ**	**INFO**	**READ**	**INFO**	**READ**
Better	100%	30%	100%	90%	100%	90%
Equal	0%	50%	0%	0%	0%	0%
Worse	0%	20%	0%	10%	0%	10%
Opinion	**Annotator 1**		**Annotator 2**		**Annotator 3**	
MSA	**INFO**	**READ**	**INFO**	**READ**	**INFO**	**READ**
Better	90%	20%	60%	70%	100%	100%
Equal	10%	80%	40%	30%	0%	0%
Worse	0%	0%	0%	0%	0%	0%

Table 16. MSA vs. BA Evaluation

6 Conclusion and Future Work

In this paper, we have carried out a comprehensive analysis of the question types in community-based question answering (cQA) services and have developed taxonomies for questions and answers. We find that questions do not always have unique best answers. *Open* and *opinion* questions usually have multiple good answers. They occupied about 56%~83% and most of their best answers can be improved. By using question type as a guide, we propose applying automatic summarization techniques to summarization answers or improving cQA best answers through answer editing. Our results show that customized question-type focused summarization techniques can improve cQA answer quality significantly.

Looking into the future, we are to develop automatic question type identification methods to fully automate answer summarization. Furthermore, we would also like to utilize more sophisticated summarization techniques to improve content compaction and readability.

Acknowledgements

We thank the anonymous reviewers for their valuable suggestions and comments to this paper.

References

Broder A. A taxonomy of web search. 2002. SIGIR Forum Vol.36, No. 2, 3-10.

Hovy Edward, Laurie Gerber, Ulf Hermjakob, Chin-Yew Lin, Deepak Ravichandran. 2001. Toward Semantics-Based Answer Pinpointing. In *Proc. of HLT'01*.

Hovy E., C. Lin. 1999. Automated Text Summarization and the SUMMARIST System. In Advances in Automated Text Summarization

Jeon J., W. B. Croft, and J. Lee. 2005a. Finding semantically similar questions based on their answers. In *Proc. of SIGIR'05*.

Jeon J., W. B. Croft, and J. Lee. 2005b. Finding similar questions in large question and answer archives. In *Proc. of CIKM'05.*

Jurczyk P., E. Agichtein. 2007. Hits on question answer portals: exploration of link analysis for author ranking. In *Proc. of SIGIR '07*.

Jeon J. , W.B. Croft, J. Lee, S. Park. 2006. A Framework to predict the quality of answers with non-textual features. In *Proc. of SIGIR '06*.

Jijkoun V., M. R. 2005. Retrieving Answers from Frequently Asked Questions Pages on the Web. In *Proc. of CIKM'05*.

Kleinberg J. 1999. Authoritative sources in a hyperlinked environment. Journal of the ACM, vol. 46,

Kim S., E. Hovy. 2004. Determining the Sentiment of Opinions. In *Proc. of COLING'04*.

Liu X., W.B. Croft, M. Koll. 2005. Finding experts in community-based question-answering services. In *Proc. of CIKM '05*.

Lin C.Y., E. Hovy. 2002. From single to multi-document summarization: a prototype system and its evaluation. In *Proc. of ACL'02*.

Lytinen S., N. Tomuro. 2002. The Use of Question Types to Match Questions in FAQFinder. In *Proc. of AAAI'02.*

Moldovan D., S. Harabagiu, et al. 2000. The Structure and an Open-Domain Question Answering System. In *Proc. of ACL'00*.

Mei Q., X. Shen, C. Zhai. 2007. Automatic labeling of multinomial topic models. In *Proc. of KDD'07*.

Rose D. E., D. Levinson. 2004. Understanding user goals in web search. In *Proc. of WWW '04*.

Voorhees, M. Ellen. 2003. Overview of the TREC 2003 Question Answering Track. In *Proc. of TREC'03.*

Wiebe J., E. Breck, et al. 2003. Recognizing and Organizing Opinions Expressed in the World Press