# Community Answer Summarization for Multi-Sentence Question with Group $L_1$ Regularization

**Wen Chan†, Xiangdong Zhou†, Wei Wang†, Tat-Seng Chua‡**
†School of Computer Science, Fudan University, Shanghai, 200433, China
{11110240007,xdzhou,weiwang1}@fudan.edu.cn
‡School of Computing, National University of Singapore
chuats@nus.edu.sg

## Abstract

We present a novel answer summarization method for community Question Answering services (cQAs) to address the problem of "incomplete answer", i.e., the "best answer" of a complex multi-sentence question misses valuable information that is contained in other answers. In order to automatically generate a novel and non-redundant community answer summary, we segment the complex original multi-sentence question into several sub questions and then propose a general Conditional Random Field (CRF) based answer summary method with group $L_1$ regularization. Various textual and non-textual QA features are explored. Specifically, we explore four different types of contextual factors, namely, the information novelty and non-redundancy modeling for local and non-local sentence interactions under question segmentation. To further unleash the potential of the abundant cQA features, we introduce the group $L_1$ regularization for feature learning. Experimental results on a Yahoo! Answers dataset show that our proposed method significantly outperforms state-of-the-art methods on cQA summarization task.

## 1 Introduction

Community Question and Answering services (cQAs) have become valuable resources for users to pose questions of their interests and share their knowledge by providing answers to questions. They perform much better than the traditional frequently asked questions (FAQ) systems (Jijkoun and Rijke , 2005; Riezler et al., 2007) which are just based on natural language processing and information retrieving technologies due to the need for human intelligence in user generated contents(Gyongyi et al., 2007). In cQAs such as Yahoo! Answers, a resolved question often gets more than one answers and a "best answer" will be chosen by the asker or voted by other community participants. This {question, best answer} pair is then stored and indexed for further uses such as question retrieval. It performs very well in simple factoid QA settings, where the answers to factoid questions often relate to a single named entity like a person, time or location. However, when it comes to the more sophisticated multi-sentence questions, it would suffer from the problem of "incomplete answer". That is, such question often comprises several sub questions in specific contexts and the asker wishes to get elaborated answers for as many aspects of the question as possible. In which case, the single best answer that covers just one or few aspects may not be a good choice (Liu et al., 2008; Takechi et al., 2007). Since "everyone knows something" (Adamic et al., 2008), the use of a single best answer often misses valuable human generated information contained in other answers.

In an early literature, Liu et al.(2008) reported that no more than $48\%$ of the 400 best answers were indeed the unique best answers in 4 most popular Yahoo! Answers categories. Table 1 shows an example of the "incomplete answer" problem from Yahoo! Answers[1]. The asker wishes to know why his teeth bloods and how to prevent it. However, the best answer only gives information on the reason of teeth

[1] http://answers.yahoo.com/question/index?qid=20100610161858AAmAGrV

*Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 582–591,
Jeju, Republic of Korea, 8-14 July 2012. ©2012 Association for Computational Linguistics

blooding. It is clear that some valuable information about the reasons of gums blooding and some solutions are presented in other answers.

| |
|---|
| **Question** |
| Why do teeth bleed at night and how do you prevent/stop it? This morning I woke up with blood caked between my two front teeth. This is the third morning in a row that it has happened. I brush and floss regularly, and I also eat a balanced, healthy diet. Why is this happening and how do I stop it? |
| **Best Answer - Chosen by Asker** |
| Periodontal disease is a possibility, gingivitis, or some gum infection. Teeth don't bleed; gums bleed. |
| **Other Answers** |
| Vitamin C deficiency! |
| Ever heard of a dentist? Not all the problems in life are solved on the Internet. |
| You could be brushing or flossing too hard. Try a brush with softer bristles or brushing/flossing lighter and slower. If this doesn't solve your problem, try seeing a dentist or doctor. Gums that bleed could be a sign of a more serious issue like leukemia, an infection, gum disease, a blood disorder, or a vitamin deficiency. |
| wash your mouth with warm water and salt, it will help to strengthen your gum and teeth, also salt avoid infection. You probably have weak gums, so just try to follow the advice, it works in many cases of oral problems. |

Table 1: An example of question with incomplete answer problem from Yahoo! Answers. The "best answer" seems to miss valuable information and will not be ideal for re-use when similar question is asked again.

In general, as noted in (Jurafsky and Martin , 2009), most interesting questions are not factoid questions. User's needs require longer, more informative answers than a single phrase. In fact, it is often the case, that a complex multi-sentence question could be answered from multiple aspects by different people focusing on different sub questions. Therefore we address the incomplete answer problem by developing a novel summarization technique taking different sub questions and contexts into consideration. Specifically we want to learn a concise summary from a set of corresponding answers as supplement or replacement to the "best answer".

We tackle the answer summary task as a sequential labeling process under the general Conditional Random Fields (CRF) framework: every answer sentence in the question thread is labeled as a *summary* sentence or *non-summary* sentence, and we concatenate the sentences with *summary* label to form the final summarized answer. The contribution of this paper is two-fold:

First, we present a general CRF based framework and incorporate four different contextual factors based on question segmentation to model the local and non-local semantic sentence interactions to address the problem of redundancy and information novelty. Various textual and non-textual question answering features are exploited in the work.

Second, we propose a group $L_1$-regularization approach in the CRF model for automatic optimal feature learning to unleash the potential of the features and enhance the performance of answer summarization.

We conduct experiments on a Yahoo! Answers dataset. The experimental results show that the proposed model improve the performance significantly(in terms of precision, recall and F1 measures) as well as the ROUGE-1, ROUGE-2 and ROUGE-L measures as compared to the state-of-the-art methods, such as Support Vector Machines (SVM), Logistic Regression (LR) and Linear CRF (LCRF) (Shen et al., 2007).

The rest of the paper is arranged as follows: Section 2 presents some definitions and a brief review of related research. In Section 3, we propose the summarization framework and then in Section 4 and 5 we detail the experimental setups and results respectively. We conclude the paper in Section 6.

## 2 Definitions and Related Work

### 2.1 Definitions

In this subsection we define some concepts that would be helpful to clarify our problems. First we define a *complex multi-sentence* question as a question with the following properties:

**Definition:** A ***complex multi-sentence question*** is one that contains multiple sub-questions.

In the cQAs scenario a question often consists of one or more main question sentences accompany by some context sentences described by askers. We treat the original question and context as a whole single complex multi-sentence question and obtain the sub questions by question segmentation. We then define the *incomplete answer* problem as:

**Definition:** The ***incomplete answer problem*** is one where the best answer of a complex multi-sentence question is voted to be below certain star ratings or the average similarity between the best answer and all the sub questions is below some thresholds.

We study the issues of similarity threshold and the minimal number of stars empirically in the experimental section and show that they are useful in identifying questions with the incomplete answer problem.

### 2.2 Related Work

There exist several attempts to alleviate the answer completeness problem in cQA. One of them is to segment the multi-sentence question into a set of sub-questions along with their contexts, then sequentially retrieve the sub questions one by one, and return similar questions and their best answers (Wang et al., 2010). This strategy works well in general, however, as the automatic question segmentation is imperfect and the matched similar questions are likely to be generated in different contextual situations, this strategy often could not combine multiple independent best answers of sub questions seamlessly and may introduce redundancy in final answer.

On general problem of cQA answer summarization, Liu et al.(2008) manually classified both questions and answers into different taxonomies and applied clustering algorithms for answer summarization.They utilized textual features for open and opinion type questions. Through exploiting metadata, Tomasoni and Huang(2010) introduced four characteristics (constraints) of summarized answer and combined them in an additional model as well as a multiplicative model. In order to leverage context, Yang et al.(2011) employed a dual wing factor graph to mutually enhance the performance of social document summarization with user generated content like tweets. Wang et al. (2011) learned online discussion structures such as the replying relationship by using the general CRFs and presented a detailed description of their feature designs for sites and edges embedded in discussion thread structures. However there is no previous work that explores the complex multi-sentence question segmentation and its contextual modeling for community answer summarization.

Some other works examined the evaluation of the quality of features for answers extracted from cQA services (Jeon et al., 2006; Hong and Davison , 2009; Shah et al., 2010). In the work of Shah et al.(2010), a large number of features extracted for predicting asker-rated quality of answers was evaluated by using a logistic regression model. However, to the best of our knowledge, there is no work in evaluating the quality of features for community answer summarization. In our work we model the feature learning and evaluation problem as a group $L_1$ regularization problem (Schmidt , 2010) on different feature groups.

## 3 The Summarization Framework

### 3.1 Conditional Random Fields

We utilize the probabilistic graphical model to solve the answer summarization task, Figure 1 gives some illustrations, in which the sites correspond to the sentences and the edges are utilized to model the interactions between sentences. Specifically, let **x** be the sentence sequence to all answers within a question thread, and **y** be the corresponding label sequence. Every component $y_i$ of **y** has a binary value, with +1 for the summary sentence and -1 otherwise. Then under CRF (Lafferty et al., 2001), the conditional probability of **y** given **x** obeys the following distribution:

$$\begin{aligned} p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} exp(&\sum_{v \in V, l} \mu_l g_l(v, \mathbf{y}|_v, \mathbf{x}) \\ &+ \sum_{e \in E, k} \lambda_k f_k(e, \mathbf{y}|_e, \mathbf{x})), \end{aligned} \tag{1}$$

where $Z(\mathbf{x})$ is the normalization constant called partition function, $g_l$ denotes the cQA feature function of site $l$, $f_k$ denotes the function of edge $k$( modeling the interactions between sentences), $\mu$ and $\lambda$ are respectively the weights of function of sites and edges, and $\mathbf{y}|_t$ denotes the components of **y** related to site (edge) $t$.

### 3.2 cQA Features and Contextual Modeling

In this section, we give a detailed description of the different sentence-level cQA features and the contextual modeling between sentences used in our model for answer summarization.

**Sentence-level Features**

Different from the conventional multi-document summarization in which only the textual features are utilized, we also explore a number of non-textual author related features (Shah et al., 2010) in cQAs.

**The textual features used are**:

**1.** ***Sentence Length***: The length of the sentence in the answers with the stop words removed. It seems that a long sentence may contain more information.

**2.** ***Position***: The sentence's position within the answer. If a sentence is at the beginning or at the end of one answer, it might be a generation or viewpoint sentence and will be given higher weight in the summarization task.

**3.** ***Answer Length***: The length of the answer to which the sentence belonged, again with the stop words removed.

**4.** ***Stopwords Rate***: The rate of stop words in the sentence. If a sentence contains too many stop words, it is more likely a spam or chitchat sentence rather than an informative one.

**5.** ***Uppercase Rate***: The rate of uppercase words. Uppercase words are often people's name, address or other name entities interested by askers.

**6.** ***Has Link*** Whether the sentence contains a hyperlink or not. The link often points to a more detailed information source.

**7.** ***Similarity to Question***: Semantic similarity to the question and question context. It imports the semantic information relevance to the question and question context.

**The non-textual features used include**:

**8.** ***Best Answer Star***: The stars of the best answer received by the askers or voters.

**9.** ***Thumbs Up***: The number of thumbs-ups the answer which contains the sentence receives. Users are often used to support one answer by giving a thumbs up after reading some relevant or interesting information for their intentions.

**10.** ***Author Level***: The level of stars the author who gives the answer sentence acquires. The higher the star level, the more authoritative the asker is.

**11.** ***Best Answer Rate***: Rate of answers annotated as the best answer the author who gives the answer sentence receives.

**12.** ***Total Answer Number***: The number of total answers by the author who gives the answer sentence. The more answers one gives, the more experience he or she acquires.

**13.** ***Total Points***: The total points that the author who gives the answer sentence receives.

The previous literature (Shah et al., 2010) hinted that some cQA features, such as *Sentence Length*, *Has Link* and *Best Answer Star*, may be more important than others. We also expect that some feature may be redundant when their most related features are given, e.g., the *Author Level* feature is positively related with the *Total Points* received by answerers, and *Stopwords Rate* is of little help when both *Sentence Length* (not including stop words) and U*ppercase Rate* are given. Therefore, to explore the optimal combination of these features, we propose a group $L_1$ regularization term in the general CRF model (Section 3.3) for feature learning.

All features presented here can be extracted automatically from the Yahoo! Answers website. We normalize all these feature values to real numbers between 0 and 1 by dividing them by the corresponding maximal value of these features. These sentence-level features can be easily utilized in the CRF framework. For instance, if the rate of uppercase words is prominent or the position is close to the beginning or end of the answer, then the probability of the label +1 (summary sentence) should be boosted by assigning it with a large value.

### Contextual Modeling Under Question Segmentation

For cQAs summarization, the semantic interactions between different sentence sites are crucial, that is, some context co-occurrences should be encouraged and others should be penalized for requirements of information novelty and non-redundancy in the generated summary. Here we consider both local (sentences from the same answer) and global (sentences from different answers) settings. This give rise to four contextual factors that we will explore for modeling the pairwise semantic interactions based on question segmentation. In this paper, we utilize a simple but effective lightweight question segmentation method (Ding et al., 2008; Wang et al., 2010). It mainly involves the following two steps:

**Step 1**. Question sentence detection: every sentence in the original multi-sentence question is classified into *question sentence* and *non-question (context) sentence*. The *question mark* and *5W1H* features are applied.

**Step 2**. Context assignment: every context sentence is assigned to the most relevant question sentence. We compute the semantic similarity(Simpson and Crowe, 2005) between sentences or sub ques-

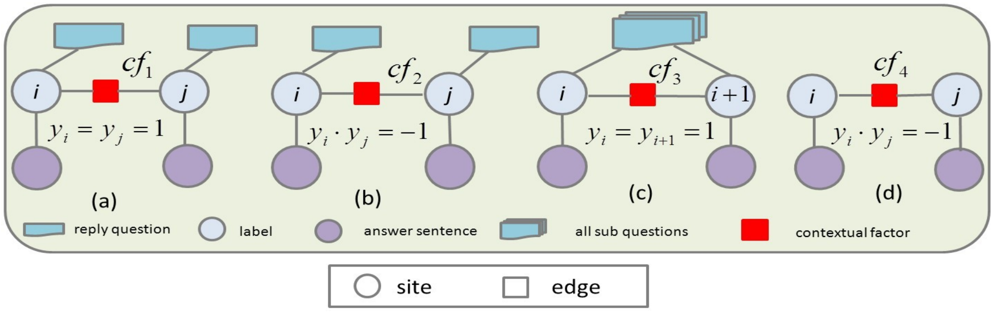

Figure 1: Four kinds of the contextual factors are considered for answer summarization in our general CRF based models.

tions as:

$$sim(x,y) = 2 \times \sum_{(w_1,w_2)\in M(x,y)} \frac{sim(w_1,w_2)}{|x|+|y|} \quad (2)$$

where $M(x,y)$ denotes synset pairs matched in sentences $x$ and $y$; and the similarity between the two synsets $w_1$ and $w_2$ is computed to be inversely proportional to the length of the path in Wordnet.

One answer sentence may related to more than one sub questions to some extent. Thus, we define the *replied question* $Qr_i$ as the sub question with the maximal similarity to sentence $x_i$: $Qr_i = argmax_{Q_j} sim(x_i, Q_j)$. It is intuitive that different summary sentences aim at answering different sub questions. Therefore, we design the following two contextual factors based on the similarity of replied questions.

**Dissimilar Replied Question Factor**: Given two answer sentences $x_i$ , $x_j$ and their corresponding replied questions $Qr_i$, $Qr_j$. If the similarity[2] of $Qr_i$ and $Qr_j$ is below some threshold $\tau_{lq}$, it means that $x_i$ and $x_j$ will present different viewpoints to answer different sub questions. In this case, it is likely that $x_i$ and $x_j$ are both summary sentences; we ensure this by setting the contextual factor $cf_1$ with a large value of $exp\ \nu$, where $\nu$ is a positive real constant often assigned to value 1; otherwise we set $cf_1$ to $exp - \nu$ for penalization.

$$cf_1 = \begin{cases} exp\ \ \nu,\ y_i = y_j = 1 \\ exp - \nu,\ otherwise \end{cases}$$

**Similar Replied Question Factor**: Given two answer sentences $x_i$ , $x_j$ and their corresponding replied questions $Qr_i$, $Qr_j$. If the similarity of $Qr_i$ and $Qr_j$ is above some upper threshold $\tau_{uq}$, this means that $x_i$ and $x_j$ are very similar and likely to provide similar viewpoint to answer similar questions. In this case, we want to select either $x_i$ or $x_j$ as answer. This is done by setting the contextual factor $cf_2$ such that $x_i$ and $x_j$ have opposite labels,

$$cf_2 = \begin{cases} exp\ \ \nu,\ y_i * y_j = -1 \\ exp\ \ -\nu,\ otherwise \end{cases}$$

Assuming that sentence $x_i$ is selected as a summary sentence, and its next local neighborhood sentence $x_{i+1}$ by the same author is dissimilar to it but it is relevant to the original multi-sentence question, then it is reasonable to also pick $x_{i+1}$ as a summary sentence because it may offer new viewpoints by the author. Meanwhile, other local and non-local sentences which are similar to it at above the upper threshold will probably not be selected as summary sentences as they offer similar viewpoint as discussed above. Therefore, we propose the following two kinds of contextual factors for selecting the answer sentences in the CRF model.

**Local Novelty Factor**: If the similarity of answer sentence $x_i$ and $x_{i+1}$ given by the same author is below a lower threshold $\tau_{ls}$, but their respective similarities to the sub questions both exceed an upper threshold $\tau_{us}$, then we will boost the probability of selecting both as summary sentences by setting:

$$cf_3 = \begin{cases} exp\ \ \nu,\ y_i = y_{i+1} = 1 \\ exp - \nu,\ otherwise \end{cases}$$

**Redundance Factor**: If the similarity of answer

[2] We use the semantic similarity of Equation 2 for all our similarity measurement in this paper.

sentence $x_i$ and $x_j$ is greater than the upper threshold $\tau_{us}$, then they are likely to be redundant and hence should be given opposite labels. This is done by setting:

$$cf_4 = \begin{cases} exp\ \nu,\ y_i * y_j = -1 \\ exp - \nu,\ otherwise \end{cases}$$

Figure 1 gives an illustration of these four contextual factors in our proposed general CRF based model. The parameter estimation and model inference are discussed in the following subsection.

### 3.3 Group $L_1$ Regularization for Feature Learning

In the context of cQA summarization task, some features are intuitively to be more important than others. As a result, we group the parameters in our CRF model with their related features[3] and introduce a group $L_1$-regularization term for selecting the most useful features from the least important ones, where the regularization term becomes,

$$R(\theta) = C \sum_{g=1}^{G} \|\overrightarrow{\theta_g}\|_2, \qquad (3)$$

where $C$ controls the penalty magnitude of the parameters, $G$ is the number of feature groups and $\overrightarrow{\theta_g}$ denotes the parameters corresponding to the particular group $g$. Notice that this penalty term is indeed a $L(1,2)$ regularization because in every particular group we normalize the parameters in $L_2$ norm while the weight of a whole group is summed in $L_1$ form.

Given a set of training data $D = (x^{(i)}, y^{(i)})$, $i = 1, ..., N$, the parameters $\theta = (\mu_l, \lambda_k)$ of the general CRF with the group $L_1$-regularization are estimated in using a maximum log likelihood function $L$ as:

$$L = \sum_{i=1}^{N} log(p_\theta(y^{(i)}|x^{(i)})) - C \sum_{g=1}^{G} \|\overrightarrow{\theta_g}\|_2, \qquad (4)$$

where $N$ denotes the total number of training samples. we compute the log-likelihood gradient component of $\theta$ in the first term of Equation 4 as in usual CRFs. However, the second term of Equation 4 is non-differentiable when some special $\|\overrightarrow{\theta_g}\|_2$ becomes exactly zero. To tackle this problem, an additional variable is added for each group (Schmidt , 2010); that is, by replacing each norm $\|\overrightarrow{\theta_g}\|_2$ with the variable $\alpha_g$, subject to the constraint $\alpha_g \geq \|\overrightarrow{\theta_g}\|_2$, i.e.,

$$L = \sum_{i=1}^{N} log(p_\theta(y^{(i)}|x^{(i)})) - C \sum_{g=1}^{G} \alpha_g, \quad subject\ to\ \alpha_g \geq \|\overrightarrow{\theta_g}\|_2, \forall g. \qquad (5)$$

This formulation transforms the non-differentiable regularizer to a simple linear function and maximizing Equation 5 will lead to a solution to Equation 4 because it is a lower bound of the latter. Then, we add a sufficient small positive constant $\varepsilon$ when computing the $L_2$ norm (Lee et al., 2006), i.e., $|\overrightarrow{\theta_g}\|_2 = \sqrt{\sum_{j=1}^{|g|} \theta_{gj}^2 + \varepsilon}$, where $|g|$ denotes the number of features in group $g$. To obtain the optimal value of parameter $\theta$ from the training data, we use an efficient L-BFGS solver to solve the problem, and the first derivative of every feature $j$ in group $g$ is,

$$\frac{\delta L}{\delta \theta_{gj}} = \sum_{i=1}^{N} C_{gj}(y^{(i)}, x^{(i)}) - \sum_{i=1}^{N} \sum_{y} p(y|x^{(i)}) C_{gj}(y, x^{(i)}) - 2C \frac{\theta_{gj}}{\sqrt{\sum_{l=1}^{|g|} \theta_{gl}^2 + \varepsilon}} \qquad (6)$$

where $C_{gj}(y, x)$ denotes the count of feature $j$ in group $g$ of observation-label pair $(x, y)$. The first two terms of Equation 6 measure the difference between the empirical and the model expected values of feature $j$ in group $g$, while the third term is the derivative of group $L_1$ priors.

For inference, the labeling sequence can be obtained by maximizing the probability of $y$ conditioned on $x$,

$$y^* = argmax_y p_\theta(y|x). \qquad (7)$$

We use a modification of the Viterbi algorithm to perform inference of the CRF with non-local edges

[3] We note that every sentence-level feature discussed in Section 3.2 presents a variety of instances (e.g., the sentence with longer or shorter length is the different instance), and we may call it sub-feature of the original sentence-level feature in the micro view. Every sub-feature has its corresponding weight in our CRF model. Whereas in a macro view, those related sub-features can be considered as a group.

previously used in (Galley , 2006). That is , we replace the edge connection $z_t = (y_{t-2}, y_{t-1}, y_t)$ of order-2 Markov model by $z_t = (y_{N_t}, y_{t-1}, y_t)$, where $y_{N_t}$ represents the label at the source of the non-local edge. Although it is an approximation of the exact inference, we will see that it works well for our answer summarization task in the experiments.

## 4 Experimental Setting

### 4.1 Dataset

To evaluate the performance of our CRF based answer summarization model, we conduct experiments on the Yahoo! Answers archives dataset. The Yahoo! $Webscope^{TM}$ Program[4] opens up a number of Yahoo! Answers datasets for interested academics in different categories. Our original dataset contains 1,300,559 questions and 2,770,896 answers in ten taxonomies from Yahoo! Answers. After filtering the questions which have less than 5 answers and some trivial factoid questions using the features by (Tomasoni and Huang, 2010) , we reduce the dataset to 55,132 questions. From this sub-set, we next select the questions with incomplete answers as defined in Section 2.1. Specifically, we select the questions where the average similarity between the best answer and all sub questions is less than 0.6 or when the star rating of the best answer is less than 4. We obtain 7,784 questions after this step. To evaluate the effectiveness of this method, we randomly choose 400 questions in the filtered dataset and invite 10 graduate candidate students (not in NLP research field) to verify whether a question suffers from the incomplete answer problem. We divide the students into five groups of two each. We consider the questions as the "incomplete answer questions" only when they are judged by both members in a group to be the case. As a result, we find that 360 (90%) of these questions indeed suffer from the incomplete answer problem, which indicates that our automatic detection method is efficient. This randomly selected 400 questions along with their 2559 answers are then further manually summarized for evaluation of automatically generated answer summaries by our model in experiments.

### 4.2 Evaluation Measures

When taking the summarization as a sequential biclassification problem, we can make use of the usual precision, recall and F1 measures (Shen et al., 2007) for classification accuracy evaluation.

In our experiments, we also compare the precision, recall and F1 score in the ROUGE-1, ROUGE-2 and ROUGE-L measures (Lin , 2004) for answer summarization performance.

## 5 Experimental Results

### 5.1 Summarization Results

We adapt the Support Vector Machine (SVM) and Logistic Regression (LR) which have been reported to be effective for classification and the Linear CRF (LCRF) which is used to summarize ordinary text documents in (Shen et al., 2007) as baselines for comparison. To better illustrate the effectiveness of question segmentation based contextual factors and the group $L_1$ regularization term, we carry the tests in the following sequence: (a) we use only the contextual factors $cf_3$ and $cf_4$ with default $L_2$ regularization (gCRF); (b) we add the reply question based factors $cf_1$ and $cf_2$ to the model (gCRF-QS); and (c) we replace default $L_2$ regularization with our proposed group $L_1$ regularization term (gCRF-QS-l1). For linear CRF system, we use all our textual and non-textual features as well as the local (exact previous and next) neighborhood contextual factors instead of the features of (Shen et al., 2007) for fairness. For the thresholds used in the contextual factors, we enforce $\tau_{lq}$ to be equal to $\tau_{ls}$ and $\tau_{uq}$ equal to $\tau_{us}$ for the purpose of simplifying the parameters setting ($\tau_{lq} = \tau_{ls} = 0.4$, $\tau_{uq} = \tau_{us} = 0.8$ in our experiments). We randomly divide the dataset into ten subsets (every subset with 40 questions and the associated answers), and conduct a ten-fold cross validation and for each round where the nine subsets are used to train the model and the remaining one for testing. The precision, recall and F1 measures of these models are presented in Table 2.

Table 2 shows that our general CRF model based on question segmentation with group $L_1$ regularization out-performs the baselines significantly in all three measures (gCRF-QS-l1 is 13.99% better than SVM in precision, 9.77% better in recall and 11.72% better in F1 score). We note that both SVM and LR,

[4] http://sandbox.yahoo.com/

| Model | R1_P | R1_R | R1_F1 | R2_P | R2_R | R2_F1 | RL_P | RL_R | RL_F1 |
|---|---|---|---|---|---|---|---|---|---|
| SVM | 79.2% | 52.5% | 63.1% | 71.9% | 41.3% | 52.4% | 67.1% | 36.7% | 47.4% |
| LR | 75.2%↓ | 57.4%↑ | 65.1%↑ | 66.1%↓ | 48.5%↑ | 56.0%↑ | 61.6%↓ | 43.2%↑ | 50.8%↑ |
| LCRF | 78.7%- | 61.8%↑ | 69.3%- | 71.4%- | 54.1%↑ | 61.6%↑ | 67.1%- | 49.6%↑ | 57.0%↑ |
| ***gCRF*** | 81.9%↑ | 65.2%↑ | 72.6%↑ | 76.8%↑ | 57.3%↑ | 65.7%↑ | 73.9%↑ | 53.5%↑ | 62.1%↑ |
| ***gCRF-QS*** | 81.4%- | **70.0%**↑ | 75.3%↑ | 76.2%- | **62.4%**↑ | 68.6%↑ | 73.3%- | **58.6%**↑ | 65.1%↑ |
| ***gCRF-QS-l1*** | **86.6%**↑ | 68.3%- | **76.4%**↑ | **82.6%**↑ | 61.5%- | **70.5%**↑ | **80.4%**↑ | 58.2%- | **67.5%**↑ |

Table 3: The Precision, Recall and F1 of ROUGE-1, ROUGE-2, ROUGE-L in the baselines SVM,LR, LCRF and our general CRF based models (gCRF, gCRF-QS, gCRF-QS-l1). The down-arrow means performance degradation with statistical significance.

| Model | Precision | Recall | F1 |
|---|---|---|---|
| SVM | 65.93% | 61.96% | 63.88% |
| LR | 66.92%- | 61.31%- | 63.99%- |
| LCRF | 69.80% ↑ | 63.91%- | 66.73%↑ |
| ***gCRF*** | 73.77%↑ | 69.43%↑ | 71.53%↑ |
| ***gCRF-QS*** | 74.78%↑ | **72.51%**↑ | 73.63%↑ |
| ***gCRF-QS-l1*** | **79.92%**↑ | 71.73%- | **75.60%**↑ |

Table 2: The Precision, Recall and F1 measures of the baselines SVM,LR, LCRF and our general CRF based models (gCRF, gCRF-QS, gCRF-QS-l1). The up-arrow denotes the performance improvement compared to the precious method (above) with statistical significance under p value of 0.05, the short line '-' denotes there is no difference in statistical significance.

which just utilize the independent sentence-level features, behave not vary well here, and there is no statistically significant performance difference between them. We also find that LCRF which utilizes the local context information between sentences perform better than the LR method in precision and F1 with statistical significance. While we consider the general local and non-local contextual factor $cf_3$ and $cf_4$ for novelty and non-redundancy constraints, the gCRF performs much better than LCRF in all three measures; and we obtain further performance improvement by adding the contextual factors based on QS, especially in the recall measurement. This is mainly because we have divided the question into several sub questions, and the system is able to select more novel sentences than just treating the original multi-sentence as a whole. In addition, when we replace the default $L_2$ regularization by the group $L_1$ regularization for more efficient feature weight learning, we obtain a much better performance in precision while not sacrificing the recall measurement statistically.

We also compute the Precision, Recall and F1 in ROUGE-1, ROUGE-2 and ROUGE-L measurements, which are widely used to measure the quality of automatic text summarization. The experimental results are listed in Table 3. All results in the Table are the average of the ten-fold cross validation experiments on our dataset.

It is observed that our gCRF-QS-l1 model improves the performance in terms of precision, recall and F1 score on all three measurements of ROUGE-1, ROUGE-2 and ROUGE-L by a significant margin compared to other baselines due to the use of local and non-local contextual factors and factors based on QS with group $L_1$ regularization. Since the ROUGE measures care more about the recall and precision of N-grams as well as common substrings to the reference summary rather than the whole sentence, they offer a better measurement in modeling the user's information needs. Therefore, the improvements in these measures are more encouraging than those of the average classification accuracy for answer summarization.

From the viewpoint of ROUGE measures we observe that our question segmentation method can enhance the recall of the summaries significantly due to the more fine-grained modeling of sub questions. We also find that the precision of the group $L_1$ regularization is much better than that of the default $L_2$ regularization while not hurting the recall significantly. In general, the experimental results show that our proposed method is more effective than other baselines in answer summarization for addressing the incomplete answer problem in cQAs.

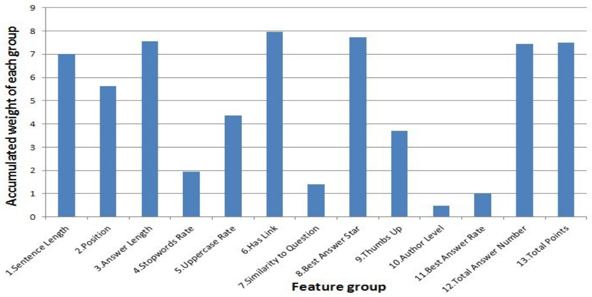

Figure 2: The accumulated weight of each site feature group in the group $L_1$-regularization to our Yahoo! Answer dataset. The horizonal axis corresponds to the name of each feature group.

### 5.2 Evaluation of Feature Learning

For group $L_1$ regularization term, we set the $\varepsilon = 10^{-4}$ in Equation 6. To see how much the different textual and non-textual features contribute to community answer summarization, the accumulated weight of each group of sentence-level features[5] is presented in Figure 2. It shows that the textual features such as 1 (*Sentence Length*), 2 (*Position*) 3 (*Answer Length*), 6 (*Has Link*) and non-textual features such as 8 (*Best Answer Star*) , 12 (*Total Answer Number*) as well as 13 (*Total Points*) have larger weights, which play a significant role in the summarization task as we intuitively considered; features 4 (*Stopwords Rate*), 5 (*Uppercase Rate*) and 9 (*Thumbs Up*) have medium weights relatively; and the other features like 7 (*Similarity to Question*), 10 (*Author Level*) and 11 (*Best Answer Rate*) have the smallest accumulated weights. The main reasons that the feature 7 (Similarity to Question) has low contribution is that we have utilized the similarity to question in the contextual factors, and this similarity feature in the single site becomes redundant. Similarly, the features *Author Level* and *Best Answer Number* are likely to be redundant when other non-textual features(*Total Answer Number* and *Total Points*) are presented together. The experimental results demonstrate that with the use of group $L_1$-regularization we have learnt better combination of these features.

### 5.3 An Example of Summarized Answer

To demonstrate the effectiveness of our proposed method, Table 4 shows the generated summary of the example question which is previously illustrated in Table 1 in the introduction section. The best answer available in the system and the summarized answer generated by our model are compared in Table 4. It is found that the summarized answer contains more valuable information about the original multi-sentence question, as it better answers the reason of teeth blooding and offers some solution for it. Storing and indexing this summarized answer in question archives should provide a better choice for answer reuse in question retrieval of cQAs.

| **Question** |
|---|
| Why do teeth bleed at night and how do you prevent/stop it? This morning I woke up with blood caked between my two front teeth.[...] |
| **Best Answer - Chosen by Asker** |
| Periodontal disease is a possibility, gingivitis, or some gum infection. Teeth don't bleed; gums bleed. |
| **Summarized Answer Generated by Our Method** |
| Periodontal disease is a possibility, gingivitis, or some gum infection. Teeth don't bleed; gums bleed. Gums that bleed could be a sign of a more serious issue like leukemia, an infection, gum disease, a blood disorder, or a vitamin deficiency. wash your mouth with warm water and salt, it will help to strengthen your gum and teeth, also salt avoid infection. |

Table 4: Summarized answer by our general CRF based model for the question in Table 1.

## 6 Conclusions

We proposed a general CRF based community answer summarization method to deal with the incomplete answer problem for deep understanding of complex multi-sentence questions. Our main contributions are that we proposed a systematic way for modeling semantic contextual interactions between the answer sentences based on question segmentation and we explored both the textual and non-textual answer features learned via a group $L_1$ regularization. We showed that our method is able to achieve significant improvements in performance of answer summarization compared to other baselines and previous methods on Yahoo! Answers dataset. We planed to extend our proposed model with more advanced feature learning as well as enriching our summarized answer with more available Web re-

[5] Note that we have already evaluated the contribution of the contextual factors in Section 5.1.

sources.

## Acknowledgements

This work was supported by the NSFC under Grant No.61073002 and No.60773077.

## References

L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman. 2008. Knowledge sharing and yahoo answers: everyone knows something. *Proceedings of WWW 2008.*

Shilin Ding, Gao Cong, Chin-Yew Lin and Xiaoyan Zhu. 2008. Rouge: Using Conditional Random Fields to Extract Contexts and Answers of Questions from Online Forums. *Proceedings of ACL-08: HLT*, pages 710–718.

Michel Galley. 2006. A Skip-Chain Conditional Random Field for Ranking Meeting Utterances by Importance. *Proceedings of EMNLP 2006.*

Z. Gyongyi, G. Koutrika, J. Pedersen, and H. Garcia-Molina. 2007. Questioning yahoo! answers. Technical report. *Stanford InfoLab.*

F. Maxwell Harper, Daphne Raban, Sheizaf Rafaeli, and Joseph A. Konstan. 2008. Predictors of Answer Quality in Online Q&A Sites. *Proceedings of CHI 2008.*

Liangjie Hong and Brian D. Davison. 2009. A Classification-based Approach to Question Answering in Discussion Boards. *Proceedings of the 32th ACM SIGIR Conference*, pages 171–178.

Eduard Hovy, Chin Y. Lin, and Liang Zhou. 2005. A BE-based Multi-document Summarization with Sentence Compression. *Proceedings of Multilingual Summarization Evaluation (ACL 2005 workshop).*

Jiwoon Jeon, W. Bruce Croft, Joon Ho Lee and Soyeon Park 2006. A Framework to Predict the Quality of Answers with NonTextual Features. *Proceedings of the 29th ACM SIGIR Conference*, pages 228–235.

V. Jijkoun and M. de Rijke. 2005. Retrieving answers from frequently asked questions pages on the web. *In CIKM.*

Daniel Jurafsky and James H. Martin. 2009. Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition. *Published by Pearson Education.*

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the 18th ICML*, pages 282–289.

S. Lee, H. Lee, P. Abbeel, and A. Ng. 2006. Efficient L1 Regularized Logistic Regression. *In AAAI.*

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Proceedings of ACL Workshop*, pages 74–81.

Yandong Liu, Jiang Bian, and Eugene Agichtein. 2008. Predicting Information Seeker Satisfaction in Community Question Answering. *Proceedings of the 31th ACM SIGIR Conference.*

Yuanjie Liu, Shasha Li, Yunbo Cao, Chin-Yew Lin, Dingyi Han, and Yong Yu. 2008. Understanding and summarizing answers in community-based question answering services. *Proceedings of the 22nd ICCL*, pages 497–504.

S. Riezler, A. Vasserman, I. Tsochantaridis, V. Mittal, and Y. Liu. 2007. Statistical machine translation for query expansion in answer retrieval. *Proceedings of the 45th Annual Meeting of ACL.*

Mark Schmidt. 2010. Graphical Model Structure Learning with L1-Regularization. *Doctoral Thesis.*

Chirag Shah and Jefferey Pomerantz. 2010. Evaluating and Predicting Answer Quality in Community QA. *Proceedings of the 33th ACM SIGIR Conference.*

Dou Shen, Jian-Tao Sun, Hua Li, Qiang Yang and Zheng Chen. 2007. Document Summarization using Conditional Random Fields. *Proceedings of the 20th IJCAI.*

Troy Simpson and Malcolm Crowe 2005. WordNet.Net http://opensource.ebswift.com/WordNet.Net

Mineki Takechi, Takenobu Tokunaga, and Yuji Matsumoto. 2007. Chunking-based Question Type Identification for Multi-Sentence Queries. *Proceedings of SIGIR 2007 Workshop.*

Mattia Tomasoni and Minlie Huang. 2010. Metadata-Aware Measures for Answer Summarization in Community Question Answering. *Proceedings of the 48th Annual Meeting of ACL*, pages 760–769.

Hongning Wang, Chi Wang, ChengXiang Zhai, Jiawei Han 2011. Learning Online Discussion Structures by Conditional Random Fields. *Proceedings of the 34th ACM SIGIR Conference.*

Kai Wang, Zhao-Yan Ming and Tat-Seng Chua. 2009. A Syntactic Tree Matching Approach to Finding Similar Questions in Community-based QA Services. *Proceedings of the 32th ACM SIGIR Conference.*

Kai Wang, Zhao-Yan Ming, Xia Hu and Tat-Seng Chua. 2010. Segmentation of Multi-Sentence Questions: Towards Effective Question Retrieval in cQA Services. *Proceedings of the 33th ACM SIGIR Conference*, pages 387–394.

X. Xue, J.Jeon, and W.B.Croft. 2008. Retrieval models for question and answers archives. *Proceedings of the 31th ACM SIGIR Conference.*

Zi Yang, Keke Cai, Jie Tang, Li Zhang, Zhou Su, and Juanzi Li. 2011. Social Context Summarization. *Proceedings of the 34th ACM SIGIR Conference.*

Liang Zhou, Chin Y. Lin, and Eduard Hovy. 2006. Summarizing answers for complicated questions. *Proceedings of the 5th International Conference on LREC, Genoa, Italy.*