A MULTI-DIMENSIONAL MODEL FOR ASSESSING THE QUALITY OF ANSWERS IN SOCIAL Q&A SITES

Technical Report TUD-CS-2009-0158
October, 2009

Zhemin Zhu, Delphine Bernhard and Iryna Gurevych
UKP Lab, Technische Universität Darmstadt
http://www.ukp.tu-darmstadt.de

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Ubiquitous
Knowledge
Processing

Abstract

The quality of user-generated content in the Web 2.0 dramatically varies from professional to abusive. Quality assessment is therefore a critical problem in producing, managing and retrieving contents in the Web 2.0. In this paper, we consider the task of assessing the quality of user-generated content, represented in natural language, as a cross-disciplinary application covered both by IQ (Information Quality) and NLP (Natural Language Processing). We develop a multi-dimensional model for assessing the quality of answers in social Q&A (Question & Answer) sites in the context of eLearning. Based on this model, we further connect potential NLP techniques with these dimensions to automate the quality assessment. Preliminary results show that it is a feasible task which opens a wide arena for techniques from both areas of IQ and NLP.

1 Introduction

The amount of user-generated content available on the Web is dramatically increasing and constitutes an important source of information in the age of Web 2.0. Many systems and modes of Web publishing, such as wikis, blogs, online discussion forums, social Q&A sites and product reviews have been developed to facilitate information production, sharing and retrieval by end-users on the World Wide Web. These platforms mainly contain information represented in natural language form rather than more structured formats. The publication threshold is rather low due to a lack of editorial control, since people who have minimal computer skills can publish information on the Web with little effort [1]. Finally, information is produced, accessed and measured in collaborative communities where most of the people involved are in equal positions.

These characteristics are double-edged. On the one hand, they make it much easier to publish on the Web. On the other hand, low quality information is more likely to appear on the Web. As a result, information on the Web varies dramatically in quality from professional to abusive.

In this paper, we focus on quality assessment of answers from social Q&A sites. Social Q&A sites such as Answerbag [1], Yahoo!Answers [2] or WikiAnswers [3] are platforms where users may post questions and get answers from fellow users. These platforms constitute the first instance of social media content that we have considered in a long-term research project which aims to cover all of the main Web 2.0 resources including social Q&A sites, wikis, blogs, online discussion forums and FAQs. Our work is set in the context of a broader project on Question Answering (QA) for eLearning based on social media content [2] whose goal is to build an automatic QA system targeted at learners. Quality assessment plays a critical role in this project since the answers delivered to the learners by the system should be especially accurate and readable. Figure 1 shows a simplified version of the architecture of the QA system. In this project, social Q&A sites are used as information source for the QA system [2].

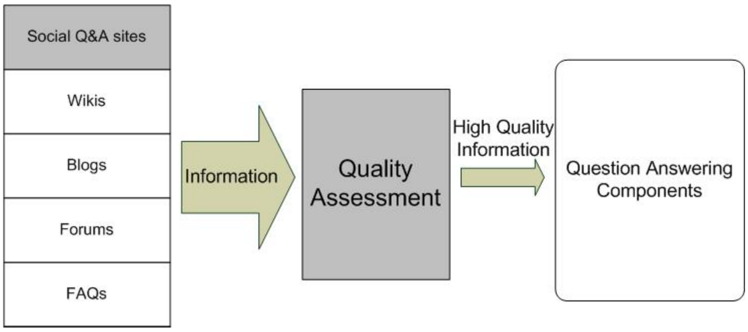

Figure 1: Architecture of our Educational Question Answering System based on Social Media Content

Two basic questions have to be answered in developing such a quality assessment component. The first question is "What is quality in the context of social Q&A sites?" and the second is "How to automatically assess quality?" Both of them are crucial. Quality is highly dependent on the type of information considered and it is therefore necessary that we provide a definition of quality with respect to social Q&A sites. Moreover, it is impossible to assess quality manually given the huge amount of information used by our QA system. We can employ techniques from the IQ (Information Quality) area to address the first question and use techniques from NLP (Natural Language Processing) to automate the quality assessment process. In this paper, we therefore define quality assessment of answers in social Q&A sites as a cross-disciplinary application covered both by IQ and NLP. The main goals of this paper are to propose a multi-

[1] http://www.answerbag.com
[2] http://answers.yahoo.com
[3] http://wiki.answers.com

dimensional model of answer quality in social Q&A sites and to connect these dimensions with the potential NLP methods to automatically assess them.

The quality model is developed gradually in three steps. The first step is dimension identification. In this step, we discover the quality related dimensions using three approaches: user survey, expert dimensions and dimensions discovered by comparing good and bad examples. Combining the dimensions extracted by these three approaches results in 13 dimensions and the corresponding metrics. The second step is dimension analysis. This aims to analyze the relations among the extracted dimensions and their relative importance in the whole quality model. Pairwise correlation, Exploratory Factor Analysis and Linear Regression Model are used. In the third step, we check the validity of the developed model for our application by applying the model to an annotated test dataset. The preliminary results are encouraging. Further, we report our ongoing work, which aims to automatically assess the dimensions obtained using NLP techniques. We describe a set of NLP techniques that may be used for this task. We found that most of the dimensions can be quantitatively assessed with the help of NLP techniques. However, a few dimensions lack established NLP techniques that can be used directly. We consider this as a motivation to develop new NLP techniques.

The article is organized as follows: related work in both IQ and NLP is described in Section 2. In Section 3, we develop the multi-dimensional quality model. In Section 4, we describe the potential NLP techniques for assessing each dimension automatically. Conclusions are given in Section 5.

2 Related Work

2.1 Related Work in IQ

Ge and Helfert [3] provide a comprehensive review of information quality research. Research in IQ follows two directions [4]: IQ management and IQ assessment. The first research direction focuses on process administration and management strategies. The second direction centers on quality [5]. In this paper, we focus on IQ assessment.

It is widely accepted that IQ is a multi-dimensional concept [5][6]. Many specific dimension-based quality models have been developed and evaluated with practical cases. Figure 2 describes a general architecture for dimension-based quality models.

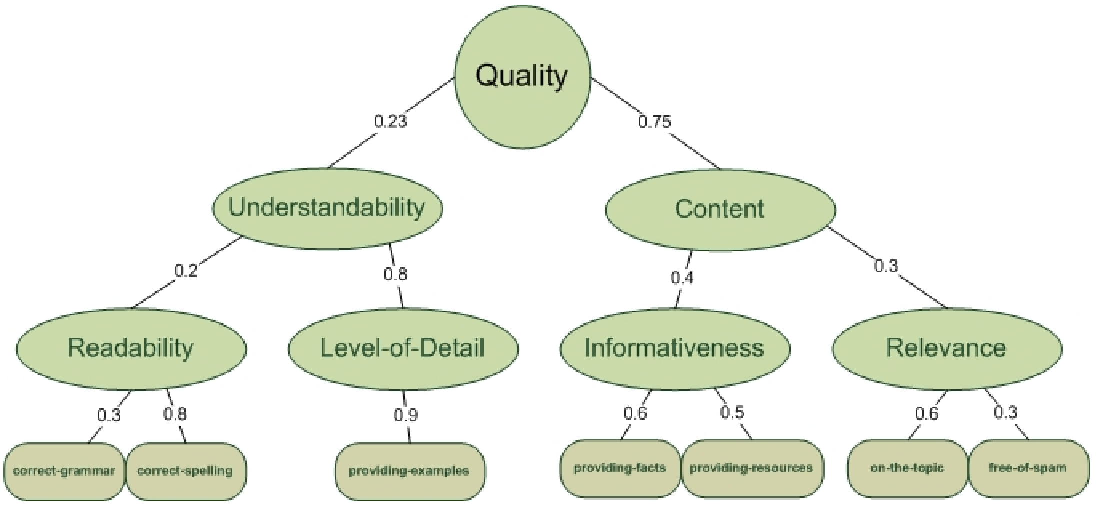

Figure 2: A dimension-based quality model

The model forms a multilevel tree structure as shown by Figure 2. The root of the tree represents overall quality. The leaves are metrics and dimensions are denoted by intermediate nodes. Metrics are usually more specific than dimensions and can normally be more easily indicated by computers or humans. The edge weights are usually calculated by some mathematic methods like Factor Analysis [7]. There are three critical tasks in developing a multi-dimensional quality model:

1. Dimension identification.

2. Dimension analysis.

3. Model validation.

The first task consists in identifying dimensions for the problem at hand. Wang and Strong [6] propose three approaches to identify dimensions:

1. The intuitive approach. This approach discovers dimensions based on the experts' experience and intuition.

2. The empirical approach. This approach derives dimensions from users of the data.

3. The comparative approach, which consists in comparing good data and bad data to discover the subtle differences between them and then identify the dimensions which can measure these differences. Wang and Wang [5] can be considered as an example of this.

Normally, the dimensions and metrics derived in the first task are not totally independent. For example, the two dimensions Free-of-Error and Understandability proposed in [8] are intuitively positively dependent on each other because error free data is usually more understandable. Therefore, relation analysis is another critical task in the development of a quality model. In the dimension identification task we define the nodes of the dimension tree as shown in Figure 2, and in the dimension analysis task we connect these nodes with edges and assign weights to these edges. The relations among dimensions can be divided into two types: (i) pairwise relations which hold between two dimensions at the same level of the dimension tree and, (ii) set relations which are among a set of dimensions. There are many mathematical tools to support analysis on these two types of relations. For the first type of relation, pairwise correlation formulations can be used like Pearson correlation. For the second type of relation, Exploratory Factorial Analysis (EFA) and clustering techniques have been employed by several papers [9][7].

When a quality model has been developed, we need to validate it for the given application. This can also be done using many mathematical methods, such as confirmation factor analysis, Cronbach's Alpha and Discriminant Validity Analysis [7].

Many dimensional quality models have been developed for specific applications. Lima et al. [7] developed a model for information in public banks in Brazil. Hammwöhner [10] and Stvilia et al. [9] build quality models for Wikipedia and use non-textual features like links and edits to analyze the dimensions. Yadav and Bellah [11] focus on the Cohesiveness dimension between Web pages to predict the quality of a website using semantic similarity.

2.2 Related Work in NLP

Quality assessment of user-generated content has also attracted a lot of interest in the NLP area. Many methods have been developed to achieve this goal. State-of-the-art NLP techniques heavily rely on machine learning algorithms. Hoang et al. [12] and Druck et al. [13] use Maximum Entropy classifier, Agichtein et al. [14] choose a log-linear model, Jeon et al. [15] adopt an in-house classifier. Liu et al. [16] and Kim et al. [17] employ the perceptron algorithm. Many other papers ([18][1][19][20][21]) choose SVMs to predict quality. Independently of the particular machine learning algorithms used in these systems, their architecture is similar with what is presented in Figure 3.

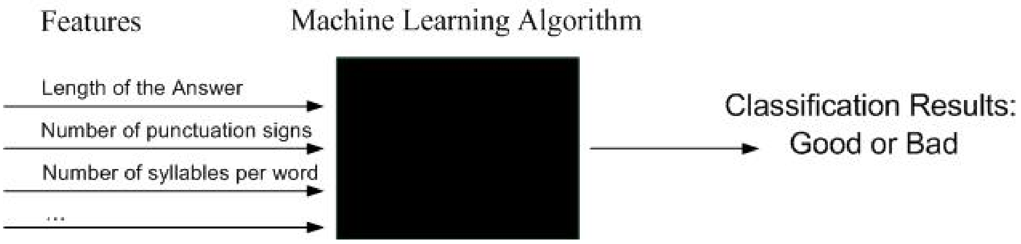
Features
Machine Learning Algorithm
Length of the Answer
Number of punctuation signs
Number of syllables per word
...
Classification Results:
Good or Bad

Figure 3: Classic Architecture for Quality Assessment in NLP

The main problem of this architecture is that the features used as input, such as length of the answer, number of punctuation signs and number of syllables per word, cannot be directly used to explain the results. The final quality judgment is not easy to interpret as the machine learning algorithm works like a black box. Intuitively, we are not inclined to accept that an answer is of high quality because it is long. Length is nevertheless a widely used feature to predict quality in NLP ([12][18][21]). This is a very shallow feature, which does not explain the overall quality judgment. However, we would rather accept that an answer is good because it is informative. In other words, existing NLP techniques fail to explain the basic question "what is quality?" in a scientifically appropriate way. This motivates us to use IQ techniques to develop a multi-dimensional model.

3 Model Development

We developed our multi-dimensional model for answers in social Q&A sites following the 3 steps presented in Section 2. The details are given in the following subsections.

3.1 Dimension Identification

To define the dimensions, we use three approaches: by a user survey, by experts' experience and intuitions, and by comparison between good and bad examples. They are detailed in the following sections.

3.1.1 Dimensions Based on a User Survey

Traditionally, making a user survey is both expensive and time-consuming. However, thanks to Web 2.0, it is easy to make an extensive survey online by using platforms such as social Q&A sites. A survey question "How do I write a good answer?"[4] was posted on Answerbag two years ago. Until now[5], it has received 185 answers, 41 comments and 476 overall votes. In Answerbag, if one person agrees or disagrees with an answer, s/he can add one point vote to the answer's total vote or subtract one point. The more votes an answer gets, the better it is. We manually extracted the dimensions and metrics from these answers and comments. Table 1 lists the results and Table 2 provides an explanation of each dimension. The following examples have been directly copied from Answerbag to illustrate how we extracted the dimensions and metrics.[6]

Example 1: "your answer should concise, easily read, do not add personal feelings to an answer." Vote : 28
We extracted the *Conciseness*, *Readability* and *Objectiveness* dimensions from this example and added 28 to the global vote of these dimensions in Table 1.

Example 2: "Don't use abbreviations, example, 'u' for the word 'you'. Don't curse, and be polite, make sure your spelling and grammar is correct. Follow all these rules and you should have a great answer." Vote: 79
For this example, we extracted the *Readability* and *Politeness* dimensions. We also identified the metrics free-of-abbreviations, correct-spelling and correct-grammar for the *Readability dimension*. The votes for these dimensions are increased by 79.

Example 3: "By being clear and providing facts or resource or personal experience that illustrates your answer." Vote: 35
Readability and *Informativeness* are extracted as dimensions and providing-facts, providing-resources, providing-personal-experience are extracted as metrics for *Informativeness*.

There are some interesting aspects in the extracted dimensions and metrics:

1. *Long answers vs. Short answers.* Some users say longer answers are better than shorter answers, but others say "keep it short". This seems to be a dilemma. In fact, this remark concerns two different dimensions. When people say that a long answer is better, they are often talking about the *Informativeness* dimension. On the other hand, when they say that short is better, they mostly refer to the *Conciseness* dimension. In summary, this example shows that Quality is a multi-dimensional concept. A good answer should be a compromise of different dimensions. This is confirmed by a sentence which is copied from the user survey: "Don't be too wordy, but give more than a 3- or 4-word answer if the question really requires it."

2. *One metric for more than one dimension.* For example, showing-a-link can be relevant to two dimensions: *Informativeness* and *Truthfulness*. There are two sentences copied from the user survey exemplifying this: "Longer answers are better than shorter answers, so stop and think about what you want to say, what point you want to make, what reference links you want to add" and "show a link to give credit to the source." Another metric provide-facts is related to these two dimensions as well. This tells us that dimensions are not totally independent. They may be connected by the same metrics.

3. *Data perspective vs. User perspective.* The *Truthfulness* dimension is an intrinsic dimension based on the data. The *Credible/Feasible/Convincing* dimension is identified from the users' perspective for different question types. *Credibility* is a user perspective measurement of *Truthfulness* for answers to factoid questions which normally begin with "What" or "Who". The *Feasible* dimension is targeted at answers to questions which begin with "How", while the *Convincing* dimension is for answers to questions beginning with "Why".

[4] http://www.answerbag.com/q_view/138108

[5] As of June 28th, 2009

[6] Please note that original answers have been directly copied from Answerbag. We did no spelling or grammar correction.

Table 1: Users Dimensions and Metrics.

Dimensions	Metrics	#Answers	Votes
Informativeness	1. providing-facts/resources/ personal experience/opinions/methods 2. no-short-phrase-answers 3. long-answers 4. showing-a-link 5. providing-explanations/comments 6. giving-examples/quotes	17	93
Politeness	1. respecting-nationality/religion/politics/race 2. proper-humor 3. humility 4. free-of-bad-words	28	165
Completeness	1. answering-the-whole-question 2. proper-humor 3. humility 4. free-of-bad-words	1	79
Readability	1. free-of-abbreviations 2. correct-spelling 3. correct-grammar 4. free-of-slang 5. no-text-talk 6. providing-paragraph-summary 7. not-all-capital-letters 8. correct-punctuation 9. proper-knowledge-level 10. simple-language 11. providing-illustrations 12. emphasis-on-important-points 13. good-organization	29	217
Relevance	1. on-the-point/topic 2. directly-answering	18	51
Conciseness or Brevity	1. keeping-short 2. directly-answering 3. minimization-of-repetition-words	9	47
Truthfulness (Credible/Feasible/Convincing)	1. backing-up-the-facts 2. explicitly-distinguishing- speculations-and-facts 3. showing-a-link 4. without-ornating-the-truth 5. providing-facts	28	50
Level of Detail	1. providing-specific-examples	2	16
Originality	None	5	12
Objectivity	1. keeping-out-personal-feelings	4	44
Novelty	1. providing-new-ideas	2	4

3.1.2 Expert Dimensions

We used the guidelines on how to write a good answer in social Q&A sites as the experts' advice. The guidelines from three social Q&A systems, namely Answerbag[7], WikiAnswers[8] and Yahoo!Answers[9] are studied. We extracted the dimensions and metrics as we did in the previous subsection. The votes are calculated as the number of guidelines from these three Q&A sites from which the dimension can be extracted. Since there are three guidelines, the votes range from 1 to 3. The results are listed in Table 3.

These guidelines tend to provide general and important principles. As shown in Table 3, a new dimension named *Usefulness* is introduced. The other dimensions overlap with the user dimensions identified in the previous subsection. Moreover, the most important dimensions which have a vote of 3 are also rated with the highest votes in the user approach.

[7] http://www.answerbag.com/guideline

[8] http://wiki.answers.com/help/answering_questions#Writing_Good_Answers

[9] http://answers.yahoo.com/info/community_guidelines

Table 2: Explanations of the Dimensions

Dimensions	Explanations
Informativeness	How suitable is the amount of information provided by the answer in relation to the question
Politeness	The degree of respect for others' feelings and opinions
Completeness	How much of the question's complete answer is covered by the given answer
Readability	How easy it is to read this answer
Relevance	How close is the answer to the subject of the question
Conciseness or Brevity	How compact is the presentation of the answer
Truthfulness (Credible/Feasible/Convincing)	How trustable is the answer
Level of Detail	How suitable is the degree of granularity
Originality	How much of the answer is not copied from other resources
Objectivity	How impartial is the answer
Novelty	How innovative is the answer

Table 3: Expert Dimensions and Metrics.

Dimensions	Metrics	Votes	Explanations
Informativeness	1. sharing-sources/knowledge/ opinions/personal experiences	3	How suitable is the amount of information provided by the answer in relation to the question
Politeness	1. free-of-sexual-content 2. free-of-personal-references: residential, business, phone number and address	3	The degree of respect for others' feelings and opinions
Readability	1. correct-spelling/grammar 2. limited-use-of-slang/abbreviations/ instant messaging-type style	3	How easy to read this answer
Relevance	1. free-of-spam (promotion of a product or an unrelated website) 2. correct-category	2	How close is the answer to the subject of the question
Truthfulness (Credible/Feasible/Convincing)	1. source-citation 2. be-accurate	1	How trustable is the answer
Originality	1. no-plagiarism 2. source-citation	1	How much of the answer is not copied from other resources
Objectivity	1. neutral-point-of-view 2. source-citation	1	How impartial is the answer
Usefulness or Helpfulness	1. addressing-the-problem	2	How useful or helpful is the answer to the asker

3.1.3 Dimensions Extracted from the Comparison between Good and Bad Examples

A third approach can be used to discover more subtle dimensions that may either be difficult to extract or have been left out by the previous two approaches. We used answers from experts[10] as good examples and answers from the same topic[11] in Answerbag for comparison (these answers may not be all bad, but comparing between an expert answer and a good answer is also helpful to identify subtle dimensions). We examined 20 pairs of answers. By comparing these 20 pairs, we discovered a new dimension and metric. The vote corresponds to the number of pairs that significantly show a difference in the *Expertise* dimension.

Table 4: Additional Dimension and Metric from Comparison

Dimensions	Metrics	Votes	Explanations
Expertise	1. professional-words/terms	16/20	The probability that the answer is written by an expert

3.1.4 Combination of the Results

Finally, we put the discovered dimensions and metrics together. This results in 13 dimensions and their metrics. The final dimensions are the following: *Informativeness*, *Politeness*, *Completeness*, *Readability*, *Relevance*, *Conciseness*, *Truthfulness (Credible/Feasible/Convincing)*, *Level of Detail*, *Originality*, *Objectivity*, *Novelty*, *Usefulness* and *Expertise*.

3.2 Dimension Analysis

In this section, we analyze the relations among the dimensions identified in the last section. We analyze two types of relations: (i) pairwise relations between dimensions and (ii) relations among a set of dimensions. We use Pearson correlation and the best-fit linear regression model for these two tasks respectively. All of these experiments are done using the free data analysis software R [22]. In NLP, detailed analyses are normally based on an annotation study. To analyze the dimension relations, we first performed an annotation study of the dimensions. One dataset was gathered from Answerbag, which consists of 50 answers. We selected the 50 first answers to the question "How do I write a good answer?", which was already analyzed in Section 3.1. Two expert raters annotated each answer with yes/no on each dimension by answering a list of questions given in Table 5.

Table 5: Questions for Annotation

Dimensions	Questions
Informativeness	Does this answer provide enough information for the question?
Politeness	Is this answer offending?
Completeness	Does this answer completely answer the whole question?
Readability	Is it easy to read this answer?
Relevance	Is this answer relevant to the question?
Conciseness or Brevity	Do you feel the answer is wordy?
Truthfulness(Credible/Feasible/Convincing)	Do you believe or trust this answer?
Level of Detail	Do you need more details?
Originality	Do you feel this answer has been copied from another place?
Objectivity	Do you feel the answer is objective and impartial?
Novelty	Are there any new ideas or concepts in this answer that make you somewhat surprised?
Usefulness or Helpfulness	Is this answer useful or helpful to address the question?
Expertise	Do you think this answer has been written by an expert?
Overall Quality	Is it a good answer?

3.2.1 Inter-rater Agreement

To check the validation of this annotation, we used Overall Agreement and Kappa to measure the inter-rater agreement. Overall agreement is calculated by dividing the number of agreed instances by the total number of instances. Kappa[12] is

[10] http://en.allexperts.com/q/Trees-739/indexExp_23328.htm

[11] We gathered the answers from the topic 'trees' in Answerbag: http://www.answerbag.com/c_view/2544

[12] We use the Kappa2 function in the R 'irr' package to calculate the Kappa values.

a statistic measurement for assessing the reliability of agreement which takes into account the agreement occurring by chance. The results are reported in Table 6. Both agreement measures indicate a good agreement between the two raters, even the lowest Kappa value of 0.429 for the dimension *Expertise* can be considered moderate agreement regarding the proposed significance in [23]. This allows us to put these two annotations together to obtain an annotation of 100 instances. The following analyses were conducted on this merged annotation.

Table 6: Inter-rater Agreement

Dimensions	**Overall Agreement**	**Cohen's Kappa**	**Rater1** (Yes/No)	**Rater2** (Yes/No)
Informativeness	0.86	0.671	38/12	39/11
Politeness	1.00	NaN	0/50	0/50
Completeness	0.92	0.792	15/35	11/39
Readability	0.92	0.558	44/6	44/6
Relevance	0.94	0.935	40/10	40/10
Conciseness	0.98	0.846	4/46	3/47
Truthfulness	0.90	0.694	44/6	46/4
Level of Detail	0.90	0.800	28/22	25/25
Originality	0.98	0.898	4/44	5/45
Objectivity	0.92	0.767	38/12	41/9
Novelty	0.90	0.699	9/41	12/38
Usefulness	0.88	0.733	38/12	37/13
Expertise	0.90	0.429	37/13	33/17
Overall Quality	0.90	0.688	37/13	37/13

3.2.2 Pairwise Correlation Analysis

To assess the pairwise relations between the dimensions, we perform Correlation analysis. The binary values yes/no are transformed to the numerical values 1 and 0, respectively. We exclude the dimension *Politeness* as all of the answers in our dataset are annotated as polite. Table 7 gives the correlation for each pair of dimensions which are calculated using the cor function in R. We also tested the significance of these relations using the cor.test function in R, in which Pearson is used as a parameter. There are some interesting points in this analysis:

Table 7: Dimension Correlation Matrix. Non significant correlation values are indicated with a star (p-value < 0.05)

	Inf.	**Com.**	**Rea.**	**Rel.**	**Con.**	**Tru.**	**Det.**	**Orig.**	**Obj.**	**Nov.**	**Use.**	**Exp.**	**Qua.**
Informativeness	1.00												
Completeness	.33	1.00											
Readability	.27	. 16	1.00										
Relevance	.68	.29	.27	1.00									
Conciseness	.03*	-.28	-.11*	.07*	1.00								
Truthfulness	.40	.21	.34	.56	-.10*	1.00							
Level of Detail	.39	.54	.19	.35	-.29	.33	1.00						
Originality	.25	-.01*	-.14*	-.01*	.40*	.08*	.01*	1.00					
Objectivity	.44	.25	.15*	.42	-.15*	.43	.21	-.03*	1.00				
Novelty	-.23	-.19	.05*	-.19*	.14*	-.05*	-.14*	.02*	-.32	1.00			
Usefulness	.81	.34	.26	.84	.02*	.46	.45	.17*	.47	-.27	1.00		
Expertise	.65	.39	.20	.57	-.18*	.47	.53	.12*	.55	-.25	.68	1.00	
Quality	.79	.35	.24	.82	.02*	.45	.47	.08*	.40	-.25	.92	.66	1.00

1. *Usefulness*, *Relevance* and *Informativeness* are the three dimensions which have the highest correlation with Overall Quality. Especially, the correlation between *Usefulness* and Overall Quality is as high as 0.92.

2. *Informativeness* and *Relevance* are highly related.

3. *Conciseness* is negatively correlated with *Readability* and *Level of Detail*. This follows our intuition. *Conciseness* has a very small correlation of 0.03 with *Informativeness*, so that we can say there is no significant relation between *Conciseness* and *Informativeness*. This shows that *Conciseness* does not necessarily indicate a loss of information.

4. *Truthfulness* is highly related with *Objectivity* and *Expertise*.

5. Only *Novelty* is negatively correlated with Overall Quality, which shows new ideas are not necessarily important for a high-quality answer.

6. *Completeness* is positively correlated with *Informativeness* and *Level of Detail* with a high correlation, and negatively correlated with *Conciseness*.

7. *Level of Detail* is positively correlated with *Expertise*. This shows that experts often give more details.

3.2.3 Factor analysis

Factor analysis was employed to discover sets of highly related dimensions. All dimensions in a set are assumed to be related with an underlying factor [24]. Grouping dimensions into subsets is helpful to discover interesting phenomena. Clustering techniques also can be applied to this task [9]. This is left for further work. We use the factanal function in the R 'FAiR' package to do factor analysis.

Two dimensions are left out for factor analysis: *Politeness* and *Usefulness*. Since all of our answers are annotated as polite, the correlation matrix is singular and thus rejected by factor analysis. Moreover *Usefulness* is related with Overall Quality with very high correlation. We tested a number of factors from 1 to 11 to find out an interpretable analysis using the default rotation. The results are given in Table 8. Factor 1 is dominated by *Relevance* and *Informativeness*, we can therefore interpret this factor as content-related. *Level of Detail* and *Completeness* play important roles in Factor 2. *Conciseness* is significantly related with this factor as well. For these reasons, we interpret Factor 2 as Understandability. *Level of Detail*, *Completeness* and *Expertise* are helpful for understanding, but *Conciseness* degrades the understandability. We are surprised that *Readability* is not so important in understandability. Factor 3 is mainly about Originality. Original answers are normally concise and provide some information. Therefore, the answers written by experts mostly seem original. But copied answers are usually more readable than the original as *Originality* is negatively related with *Readability*.

Table 8: Factor Analysis

Dimensions	**Factor 1**	**Factor 2**	**Factor 3**
Informativeness	0.741	0.251	0.250
Completeness	0.260	0.557	
Readability	0.310	0.123	-0.136
Relevance	0.904		
Conciseness	0.103	-0.578	0.380
Truthfulness	0.577	0.200	
Level of Detail	0.337	0.607	
Originality			0.997
Objectivity	0.484	0.313	
Novelty	-0.194	-0.240	
Expertise	0.627	0.512	0.130
Interpretation	**Content**	**Understandability**	**Originality**

3.3 Linear Regression Model

A linear regression model was trained to check the validity of the discovered dimensions to our application. We used the lm function in R to perform a least squares linear regression [54]. We use a step algorithm to find the numeric threshold between good answers and bad answers. Briefly, we increase the numeric threshold with a step of 0.001 from 0 to 1, and select the best one, which gives the highest accuracy. The best numeric threshold is 0.471. The dimension weights obtained are listed in Table 9.

Further, we applied this model on another larger dataset which consists of 256 question and answer pairs. These pairs were randomly selected from different categories in Answerbag. Moreover these question and answer pairs cover different question types including What, Why, How and Yes/No questions. This dataset was annotated by one of the two trained annotators employed in section 3.2. Among these 256 answers, 175 answers are annotated as good and 81 as bad. On this dataset, our model achieves an accuracy of 83.98% to predict if a given answer is good or bad based on the human annotations. We also performed an error analysis on the misclassified answers. We found that most of them are on the borderline, which means even by human it is very hard to decide if they are good or bad .

This preliminary result confirms the dimensions discovered and the linear model trained are suitable to assess the overall quality. To consider the quality as multi-dimensional is a promising direction deserving more research effort.

Table 9: Linear Regression Model

Dimensions	Weights
Intercept	-0.010631
Informativeness	0.398469
Completeness	0.007873
Readability	-0.017375
Relevance	0.560065
Conciseness	0.055291
Truthfulness	-0.065036
Level of Detail	0.077833
Originality	-0.069668
Objectivity	-0.101070
Novelty	0.018886
Expertise	0.171022

4 Connecting with NLP Techniques

To automatically assess each dimension, we plan to employ NLP techniques. This section gives a brief description of potential NLP techniques that may be used for this purpose.

The highly related techniques with *Informativeness* in NLP are IE (Information Extraction) and Opinion Mining [25]. IE [26] can be used to discover Named Entities and Opinion Mining aims to discover underlying subjective opinions in the texts. *Politeness* has been extensively studied in pragmatics, which is a subfield of NLP, in the realm of the Politeness Theory [27]. In most studies, politeness has been conceptualized as strategic conflict-avoidance and face-saving strategy [28]. *Readability* is a topic studied in NLP with a long history. Many methods have been proposed and generally these methods can be divided into two groups. The first is based on simple standard indices, such as Flesch-Kincaid [29], Gunning-Fog [30] and SMOG [31]. The second kind of methods relies on machine learning algorithms [32]. Also a lot of efforts have been invested in assessing *Relevance* in the related areas of NLP and IR (Information Retrieval). These methods differ from one another in two aspects: similarity model and the techniques used for filling the lexical gap between queries and documents ([33][34][35]). The related NLP technique to *Conciseness* are Sentence Compression [36][37] and Summarization [38]. Rubin and Liddy [39] gives four types of metrics for the assessment of *Truthfulness* for blog posts. Weerkamp and Rijke [40] goes further on this topic by automating the metrics proposed by Rubin and Liddy. *Originality* may be computed using plagiarism detection methods [41]. *Objectivity* is related to subjectivity analysis methods in NLP [25] which can be used to assess whether a sentence or an article is subjective or objective. *Novelty* can be related with the novelty detection task [42] in NLP. Given a topic, the goal of this task is to discover relevant and new information among a set of documents. Kim et al. [17] combine structural features, lexical features, syntactic features, semantic features and meta-data features to predict the *Usefulness/Helpfulness* of user-generated product reviews.

The remaining dimensions *Level of Detail*, *Completeness* and *Expertise* do not correspond to any NLP technique obviously. We will therefore strive to develop new NLP methods to assess these dimensions. The metrics discovered in this paper will help to do that. Our final system will use NLP techniques to automatically assess each dimension and the linear regression model will be applied to measure the overall quality.

5 Conclusion

Quality assessment for user-generated content is an important issue in Web 2.0. We consider this as a cross-disciplinary application between IQ and NLP. In this report, we focus on assessing the quality of answers from social Q&A sites. A multi-dimensional quality model has been developed, analyzed and evaluated. This quality model consists of 13 dimensions including *Informativeness*, *Politeness*, *Completeness*, *Readability*, *Relevance*, *Conciseness*, *Truthfulness*, *Level of Detail*, *Originality*, *Objectivity*, *Novelty*, *Usefulness* and *Expertise*. As presented in Section 3, the dimensions are linked by interesting relations. The results of this research will guide the development of an automatic quality assessment system based on NLP techniques.

References

[1] 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*, pp. 125–128, 2007.

[2] I. Gurevych, D. Bernhard, K. Ignatova, and C. Toprak, Educational question answering based on social media content, in *Proceedings of the 14th International Conference on Artificial Intelligence in Education. Building learning systems that care: From knowledge Representation to affective modelling*, pp. 133–140, 2009.

[3] M. Ge and M. Helfert, A review of information quality research-develop a research agenda, in *Proceedings of the 12th International Conference on Information Quality*, 2007.

[4] P. Oliveira, F. Rodrigues, and P. Henriques, A formal definition of data quality problems, in *Proceedings of the 10th International Conference on Information Quality*, 2005.

[5] Y. Wang and R. Y. Wang, Anchoring data quality dimensions in ontological foundations, in *Communications of ACM* Vol. 39, pp. 86–85, 1996.

[6] R. Y. Wang and D. M. Strong, Beyond accuracy: what data quality means to data consumers, in *Journal of Management Information Systems* Vol. 12, pp. 5–34, 1996.

[7] L. F. R. Lima, A. C. G. Macada, and X. Koufteros, A model for information quality in the banking industry - the case of the public banks in brazil, in *Proceedings of the 12th International Conference on Information Quality*, 2007.

[8] L. L. Pipino, Y. W. Lee, and R. Y. Wang, Data quality assessment, in *Communications of the ACM* Vol. 45, pp. 211–218, 2002.

[9] B. Stvilia, M. B. Twidale, L. C. Smith, and L. Gasser, Assessing information quality of a community-based encyclopedia, in *Proceedings of the 12th International Conference on Information Quality*, 2005.

[10] R. Hammwöhner, Interlingual aspects of wikipedia's quality, in *Proceedings of the 12th International Conference on Information Quality*, 2007.

[11] S. Yadav and J. Bellah, An improved method for automatically determining webpage cohessiveness for quality information retrival from world wide web, in *Proceedings of the 12th International Conference on Information Quality*, 2007.

[12] L. Hoang, J. T. Lee, Y. I. Song, and H. C. Rim, A model for evaluating the quality of user-created documents, in *Lecture Notes in Computer Science*, pp. 496–501, 2008.

[13] G. Druck, G. Miklau, and A. McCallum, Learning to predict the quality of contributions to wikipedia, in *Technical Report WS-08-15 published by The AAAI Press*, pp. 7–13, 2008.

[14] D. D. A. G. E. Agichtein, C. Castillo and G. Mishne, Finding high-quality content in social media, in *Proceedings of Second ACM International Conference on Web Search and Data Mining*, pp. 183–194, 2008.

[15] J. Jeon, W. B. Croft, J. H. Lee, and S. Park, A framework to predict the quality of answer with non-textual features, in *Proceedings of the 29th annual international Special Interest Group on Information Retrieval*, pp. 228–235, 2006.

[16] J. Liu, Y. Cao, C. Y. Lin, Y. Huang, and M. Zhou, Low-quality product review detection in opinion summarization, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the Conference on Computational Natural Language Learning Joint Meeting*, pp. 334–342, 2007.

[17] S. M. Kim, P. Pantel, T. Chklovski, and M. Pennacchiotti, Automatically assessing review helpfulness, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 423–430, 2006.

[18] J. E. Blumenstock, Size matters: Word count as a measure of quality on wikipedia, in *Proceedings of the 17th international conference on World Wide Web*, pp. 1095–1096, 2008.

[19] M. Surdeanu, M. Ciaramita, and H. Zaragoza, Learning to rank answers on large online qa collections, in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 719–727, 2008.

[20] N. Wanas, M. El-Saban, H. Ashour, and W. Ammar, Automatic scoring of online discussion posts, in *Proceedings of the 2nd ACM Workshop on information Credibility on the Web*, pp. 19–26, 2008.

[21] M. Weimer and I. Gurevych, Predicting the perceived quality of web forum posts, in *Proceedings of the Conference on Recent Advances in Natural Language Processing*, pp. 643–648, 2007.

[22] R. D. C. Team, R: A language and environment for statistical computing, in *R Foundation for Statistical Computing, http://www.R-project.org*, 2008.

[23] J. Landis and G. G. Koch, The measurement of observer agreement for categorical data, in *Biometrics*, pp. 159–174, 1977.

[24] A. C. Rencher, *Methods of Multivariate Analysis*, 2 ed. (Wiley-Interscience, 2002).

[25] B. Pang and L. Lee, Opinion mining and sentiment analysis, in *Foundations and Trends in Information Retrieval*, pp. 1–135, 2008.

[26] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (Person Press, 2008).

[27] L. Vilkki, Politeness, face and facework: Current issues, in *SKY Journal of Linguistics*, pp. 322–332, 2006.

[28] P. Brown and C. L. Stephen, *Politeness: Some Universals in Language Usage* (Cambridge University Press, 1987).

[29] J. P. Kincaid, Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel, in *National Technical Information Service*, pp. 8–75, 1975.

[30] R. Gunning, *The technique of clear writing* (McGraw-Hill, 1952).

[31] G. H. McLaughlin, Smog grading: A new readability formula, in *Journal of Reading*, pp. 639–646, 1969.

[32] S. E. Schwarm and M. Ostendorf, Reading level assessment using support vector machines and statistical language models, in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pp. 523–530, 2005.

[33] D. Bernhard and I. Gurevych, Combining lexical semantic resources with question & answer archives for translation-based answer finding, in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 728–736, Suntec, Singapore, 2009, Association for Computational Linguistics.

[34] H. Fang, A re-examination of query expansion using lexical resources, in *Proceedings of Annual Meeting of the Association for Computational Linguistics*, pp. 139–147, 2008.

[35] T. Zesch, C. Müller, and I. Gurevych, Using wiktionary for computing semantic relatedness, in *Proceedings of AAAI Conference on Artificial Intelligence*, pp. 861–867, 2008.

[36] J. Turner and E. Charniak, Supervised and unsupervised learning for sentence compression, in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pp. 290–297, 2005.

[37] R. Mcdonald, Discriminative sentence compression with soft syntactic constraints, in *Proceedings of the 11th The European Chapter of the ACL*, pp. 297–304, 2006.

[38] I. Mani and M. T. Maybury, *Advances in Automatic Text Summarization* (The MIT Press, 1999).

[39] V. L. Rubin and E. D. Liddy, Assessing credibility of weblogs, in *Proceedings of AAAI Conference on Artificial Intelligence*, pp. 187–191, 2006.

[40] W. Weerkamp and M. de Rijke, Credibility improves topical blog post retrieval, in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 923–931, 2008.

[41] F. K. H. Maurer and B. Zaka, Plagiarism-a survey, in *Journal of Universal Computer Science* Vol. 12, pp. 1050–1084, 2006.

[42] M. Tsai, M. Hsu, and H. Chen, Similarity computation in novelty detection, in *Proceedings of the 13th Text Retrieval Conference*, 2004.

APPENDIX

Table 10: Significance of Correlation (t-test/p-value, df=98)

	Inf.	**Com.**	**Rea.**	**Rel.**	**Con.**	**Tru.**	**Det.**
Informativeness	INFINITE						
Completeness	3.4972/ 0.0007083	INFINITE					
Readability	2.7775/ 0.006565	1.62 0.108e-1	INFINITE				
Relevance	9.2518/ 5.107e-15	2.96/ 3.78e-3	2.749/ 7.12e-3	INFINITE			
Conciseness	0.29/ 7.71e-1	-2.93/ 4.17e-3	-1.0558/ 0.2936	0.6641/ 0.5082	INFINITE		
Truthfulness	4.3347/ 3.539e-05	2.1092/ 0.03747	3.5704/ 0.0005543	6.7428/ 1.085e-09	-0.9593/ 0.3398	INFINITE	
Level of Detail	4.1731/ 6.507e-05	6.3197/ 7.756e-09	1.8666/ 0.06495	3.7463/ 0.0003032	-3.0149/ 0.003273	3.4732/ 0.000767	INFINITE
Originality	2.5718/ 0.01162	-0.101/ 0.9198	-1.3579/ 0.1776	-0.0726/ 0.9423	4.3798/ 2.979e-05	0.8014/ 0.4248	0.1078/ 0.9144
Objectivity	4.8009/ 5.669e-06	2.663/ 0.009055	1.5389/ 0.1270	4.5771/ 1.384e-05	-1.458/ 0.1481	4.7218/ 7.793e-06	2.1254/ 0.03607
Novelty	-2.3143/ 0.02274	-1.9542/ 0.05353	0.5283/ 0.5985	-1.8988/ 0.06053	1.4145/ 0.1604	-0.5368/ 0.5926	-1.4117/ 0.1612
Usefulness	13.7281/ 2.2e-16	3.6056/ 0.000492	2.6382/ 0.009695	15.2561/ 2.2e-16	0.2241/ 0.8232	5.1471/ 1.363e-06	5.0043/ 2.469e-06
Expertise	8.5586/ 1.608e-13	4.1681/ 6.631e-05	2.0328/ 0.04477	6.9202/ 4.694e-10	-1.8074/ 0.07377	5.2321/ 9.527e-07	6.1716/ 1.527e-08
Quality	12.6661/ 2.2e-16	3.7151/ 0.0003379	2.506/ 0.01386	14.03/ 2.2e-16	0.1592/ 0.8738	4.952/ 3.063e-06	5.2258/ 9.786e-07
	Orig.	**Obj.**	**Nov.**	**Use.**	**Exp.**	**Quality**	
Originality	INFINITE						
Objectivity	-0.321/ 0.749	INFINITE					
Novelty	0.2409/ 0.8102	-3.3304/ 0.001223	INFINITE				
Usefulness	1.6672/ 0.09868	5.3272/ 6.364e-07	-2.7684/ 0.006736	INFINITE			
Expertise	1.182/ 0.2400	6.483/ 3.651e-09	-2.5757/ 0.01150	9.1895/ 7.105e-15	INFINITE		
Quality	0.8251/ 0.4113	4.3289/ 3.618e-05	-2.601/ 0.01073	23.4667/ 2.2e-16	8.6201/ 1.186e-13	INFINITE	