A Model for Evaluating the Quality of User-Created Documents

Linh Hoang, Jung-Tae Lee, Young-In Song, and Hae-Chang Rim

Dept. of Computer and Radio Communications Engineering
Korea University, Seoul, Korea
{linh,jtlee,song,rim}@nlp.korea.ac.kr

Abstract. In this paper, we propose a model for evaluating the quality of general user-created documents. The model is based on supervised classification approach, in which output scores are considered as quality of given document. In order to utilize both textual and non-textual attributes of documents, we incorporated a number of objectively measurable, real-valued features selected upon predefined criteria for quality. Experiments on two datasets of real world documents show that textual features are stable indicators for evaluating documents' quality. Some features are inferred to be effective for general kinds of documents.

1 Introduction

User-created documents are well known types of user-generated contents, which are produced by end-users. For example, user product reviews in shopping sites or answers in community driven Q&A are two common types of user-created documents. This has motivated us to investigate on proposing a quality evaluating model that can be applied to any common types of user-created documents.

Using a supervised classification approach, we first manually labeled experimental documents conforming to three levels of document quality, namely *good*, *fair* and *bad*. A classifier trained from annotated corpus then ranked documents according to their prediction scores. In this work, we concentrate on building a feature combination which does not depend on the type of target documents. Our proposed method empirically worked well, even though documents have been collected from independent sources.

2 Related Work

Recently, [1] studied a task similar to our work, which is specific to user-created answers. Only non-textual features, such as *click-through counts* and *user recommendation counts*, were used for predicting answer's quality. However, it has turned out that the most effective feature is *document length* (which does not refer to non-textual information), whereas the others are less contributed. This conclusion infers that non-textual information considered previously may not always be stable along time; intuitively, data sparseness may often occur for newly created documents because they would be seen less by users.

H. Li et al. (Eds.): AIRS 2008, LNCS 4993, pp. 496–501, 2008.

[3] investigated the task of predicting reviews' helpfulness that considered users' vote as ground-truth evaluation. Firstly, different classes of features are utilized to helpfulness. SVM regression then learned helpfulness function and ranked reviews according to their output scores. In this work, the *length of a review*, *product rating* and *its unigrams* were found to be most useful. However, assessing reviews' helpfulness based on users' rating ground-truth is not always reliable due to several voting biases [4].

Showing three biases of [3]'s approach, [4] presented a framework for detecting low-quality reviews. Instead of using users' vote information, the authors manually annotated a set of ground-truth according to manually predefined specification for reviews' quality. However, many selected features are directly extracted from product's attributes such as the *number of products, product features, brand names.* Such features made this approach domain restricted since they are hardly applicable to other types of user-created documents.

Limitations from prior works have suggested us to employ both textual and non-textual features in the proposed method. To widely exploit this work for almost any types of user-created documents, only general features are chosen regarding intrinsic properties of documents. Our proposed model empirically improved performance in comparison with baseline approach that utilizes only non-textual features.

3 Method

3.1 Features Categories

One of the enhancements in our approach is the combination of objectively measurements selected upon predefined classes. All experimented features are separated into four categories: *authority*, *formality*, *readability* and *subjectivity.*

Features on authority

Among four categories, *authority* is a unique category that relies on non-textual information collected by service providers. Features on *authority* indicate whether document is written by a trustworthy author or not. Some representative examples of features in this category are as follow:

- Number of documents previously written by the same writer (NDOC)
- Number of votes or scores granted by users (NVOT)

Features on formality

This feature category refers to the writing style of target document. A formal document tends to be accessible to the intended audience. Based on this observation, some of consecutive features are considered:

- Number of words in the document (NWRD)
- Number of different words in the document (DWRD)
- Number of sentences in the document (NSNT)
- The fourth root of the number of words in the document (RWRD = $\sqrt[4]{NWRD}$)
- Average length of sentences in the document (SLEN)

Features on readability
Typically, a well-organized document imparts much information to reader. With assumption that format of document contributes to its quality prediction; three described features have been chosen for experiments:

- Lexical density of the document (LXDN = DWRD/NWRD)
- Number of paragraphs in the document (NPRG)
- Average length of paragraphs in the document (PLEN)

Features on subjectivity
Subjectivity refers to opinions of authors in a document. Several following features have been defined based on simple and easy-measurable criteria:

- Ratio of positive sentences (RPST)
- Ratio of negative sentences (RNST)
- Ratio of subjective sentences regardless of positive or negative (RSST)
- Ratio of comparative sentences (RCST)

Basically, most of features in *formality* and *readability* category are similar to the ones used in Project Essay Grade [6]. *Subjectivity* category consists of opinion-based features. Using subjective and comparative languages clues [2,8], we refined a set of opinion words and phrases for each testing corpus. *Subjectivity* features have been extracted by using a simple keyword-based approach. For example, positive sentences are considered as sentences that contain at least one positive opinion word or phrase.

3.2 Quality Evaluation Model

In our proposed model, Maximum Entropy (MaxEnt) is chosen for training a classifier. The main advantage of MaxEnt is that we can easily integrate variety of relevant features since they are expressed in the form of feature functions. For later improving a retrieval system, we intend to build a statistical model of which output scores can be considered as prior information.

The underlying idea of MaxEnt indicates that without external knowledge, one should prefer the most uniform models that also satisfy any given constraints. Once we assume that assessing quality of documents is a random process that observes documents and assign them a quality label y, MaxEnt motivates to find the model p as close to the empirical probability distribution *p'* of random process as possible. Applying to our classification task, each feature is represented by a feature function $f_i(x, y) = x_{fi}$ where x_{fi} is the value of the i^{th} feature in the document x. MaxEnt then estimate expected value for each feature from training data and take this as constraint of the model distribution.

Firstly, a set of weighting parameters λ for each feature function are estimated by using Limited-Memory Variable method [5]. The model then computes the conditional probability for predicting the quality of document x by the formula:

$$p(y|x) = \frac{1}{Z(x)} exp \left[\sum_i \lambda_i f_i(x,y) \right]$$

where *p(y|x)* is the output score indicates the quality of document x and *Z(x)* is a normalization factor to ensure $\sum_y p(y|x) = 1$. We specifically use *p(y=good|x)* as a score output.

4 Experiments

4.1 Experimental Corpus

We experimented our model on two datasets of real world user-created documents. The first one consists of 1000 English reviews on Amazon website (`http://amazon.com`). Twenty products in electronics category were randomly selected for constructing corpus. For each product, we manually accumulated 50 reviews regardless of order. Other relevant information of reviews such as author's rank, users' vote, comments are also saved. Two students were asked for hand-tagging each given document as *good*, *fair* or *bad* (Table 1 shows descriptions of three-level quality on each dataset).

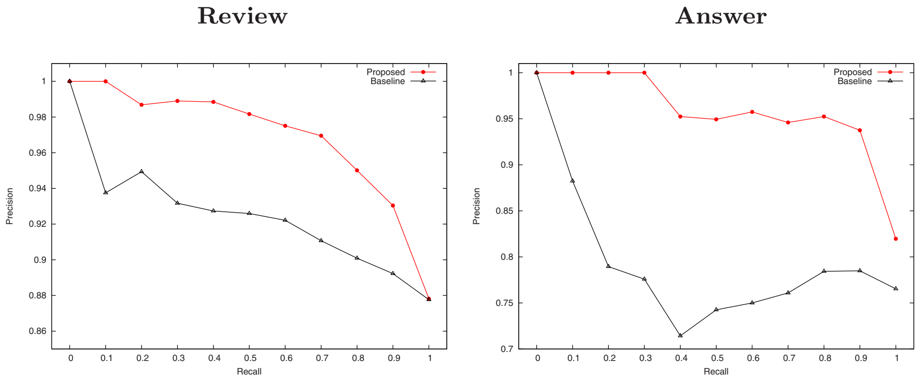
The second dataset includes 2589 Korean Q&A samples collected from Naver's Knowledge Search service (`http://kin.naver.com`). Basically, this corpus has already built and experimented in previous research [7]. The dataset is composed of questions, along with one *best* answer for each question. (Knowledge Search service allows users to select one best answer among all answers corresponding to a question). In this scenario, we used only answers for the experiments. Also, all answers were manually labeled based on three-level quality.

Table 1. Three-level specification for document quality

Level	**Document types**	
	Review	**Answer**
Good	- Complete, broad, well-organized description of the product - Pros & cons reasonably explained - Objective for most of the time	- Objective with certain basis or subjective but logically explained - Attachment often included ore answer to the question
Fair	- Contains some information about the product - Rather more subjective	- Objective but lack of details - Subjective with no basis but partially logical
Bad	- Contains very little, misleading information or even no description of the product - Many inappropriate words, wrong spellings, or bad readability - Completely subjective	- Abuse languages or spams contained - Libel on someone particular, irrelevant answer to the question - Very speculative or subjective with no basis

Table 2. Effect of feature categories

Features	**Reviews**	**Answers**
Authority(baseline)	0.7647	0.9190
+*Formality*	0.9269	0.9705
+*Readability*	0.9269	0.9674
+*Subjectivity*	0.9624	0.9722


Review

Answer

Fig. 1. 11pt recall-precision curves with 2 datasets

4.2 Results

We ranked the documents in descending order of the score output and used the traditional recall and precision metric to evaluate the results. Conforming to the evaluation metric, we consider *good* and *fair* documents as relevant documents while *poor* documents are treated as non-relevant ones. Aimed to measure the effectiveness of textual features in comparison with non-textual features, we take the model that utilizes only the features on *authority* as baseline. The average precision is chosen for measuring the overall performance and the contribution of each feature category.

Table 2 indicates the contribution of each feature category based on average precision score. From the table, *formality* is shown to be the most effective category when incorporated with non-textual features. Features on *readability* make no contribution, and even slightly decreased precision on the answer corpus. *Subjectivity* features conduced a remarkable improvement on review corpus that can be justified because of the imbalance size between two sentiment word sets.

Fig. 1 shows 11pt recall-precision curves for baseline and proposed model as well. On both of experimental datasets, textual features proved to be predictive indicator since our proposed method outperformed the baseline approach that utilizes only non-textual features.

5 Conclusion

In this paper, we presented supervised classification model for evaluating user-created documents. Four categories of features have been defined in terms of *authority*, *formality*, *readability*, *subjectivity*. We found that textual features are stable and effective for capturing document quality. Features on *formality* were pointed out to be the most useful features in augmenting the performance. *Readability* features have no significant impact, while features on *subjectivity* show promising contribution in further improvement. Although our model dominated data sparseness and restricted-domain feature selection, it still has limitations. Our proposed method only concentrated on the quality regardless of relevance to the content. Also, our experimental data for reviews is a little small.

For future work, we plan to expand our work in several practical areas such as opinion search, blog spam filtering, and summarization. We aim to continue investigating on *subjectivity* features as well as verifying the effectiveness of our quality prediction model.

Acknowledgement

This work was supported by Microsoft Research Asia. Any opinions, findings, and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsors.

References

1. Jeon, J., Croft, W.B., Lee, J.H., Park, S.: A Framework to Predict The Quality of Answers with Non-textual Features. In: SIGIR, pp. 228–235 (2006)
2. Jindal, N., Liu, B.: Identifying Comparative Sentences in Text Documents. In: SIGIR, pp. 244–251 (2006)
3. Kim, S.M., Pantel, P., Chklovski, T., Pennacchiotti, M.: Automatically Assessing Review Helpfulness. In: EMNLP, pp. 423–430 (2006)
4. Liu, J., Cao, Y., Lin, C.Y., Huang, Y., Zhou, M.: Low-Quality Product Review Detection in Opinion Summarization. In: EMNLP-CoNLL, pp. 334–342 (2007)
5. Malouf, R.: A Comparison of Algorithms for Maximum Entropy Parameter Estimation. In: CoNLL, pp. 49–55 (2002)
6. Page, E.: Computer Grading of Student Prose. JEE 62(2), 127–142 (1994)
7. Park, S., Lee, J.H., Jeon, J.: Evaluation of The Documents from The Web-based Question and Answer Service. Journal of KSLIS, 299–314 (2006)
8. Riloff, E., Wiebe, J.: Learning Extraction Patterns for Subjective Expression. In: EMNLP, pp. 105–112 (2003)