Automatic Terminology Intelligibility Estimation for Readership-oriented Technical Writing

Yasuko Senda*, **Yasusi Sinohara***, **Manabu Okumura**†

* System Engineering Research Laboratory,
Central Research Institute of Electric Power Industry, Tokyo Japan
{senda, sinohara}@criepi.denken.or.jp
† Precision and Intelligence Laboratory
Tokyo Institute of Technology, Tokyo Japan
oku@pi.titech.ac.jp

Abstract

This paper describes automatic terminology intelligibility estimation for readership-oriented technical writing. We assume that the term frequency weighted by the types of documents can be an indicator of the term intelligibility for a certain readership. From this standpoint, we analyzed the relationship between the following: average intelligibility levels of 46 technical terms that were rated by about 120 laymen; numbers of documents that an Internet search engine retrieves using each term as a keyword from various types of websites (i.e. term frequencies). The result of the analysis shows that term intelligibility for a target readership can be estimated by regression analysis of the term frequencies weighed by the type of website. As pilot studies, we developed two regression models for estimating the technical term intelligibility for the target readership. One uses the machine learning method based on ν-SVR, and the other uses multiple regression. In order to evaluate the models, we used the results of a survey on laymen's intelligibility levels for 50 new technical terms, and then compared the survey results with our estimated results. The results gave a correlation coefficient of 0.66 between the survey results and estimated results.

1. Introduction

When we write a document, we should use terms that are intelligible to the target readership. In order to avoid terms that are unintelligible to the reader, it is important to accurately evaluate term intelligibility for them. However, the more different the author's background is from the target reader's, the more difficult to objectively evaluate term intelligibility for them. Thus, if a method to estimate term intelligibility can be found, it is expected to be helpful for authors who compose documents intended for a target readership.

To solve this problem, we propose an automatic method that estimates term intelligibility and that can adjust to a target readership. In this paper, we refer to other related work in Section 2., our approaches to our research in Section 3., an overview of the sample data and evaluation data in Section 4., our proposed models and those evaluations in Section 5. and 6., and our conclusions and future work in Section 7..

2. Related Work

The several existing studies on term intelligibility are divided into two categories. One is a subjective term intelligibility rating and the other is based on the statistical term frequency.

In the former category (Amano and Kondo, 1998), an intelligibility database was developed for about 80,000 Japanese headwords taken from a commercial dictionary, for which intelligibility was rated by 43 Japanese adults on a seven-point scale through a questionnaire. However, the database has the following disadvantages:.

- The rating is valid only for the average person, and cannot be adapted for a specific readership.
- It is difficult to cover some terms, especially new terms, compound words, and technical terms because a questionnaire survey that covers a large vocabulary would be too costly to conduct.
- It becomes obsolete with time because the questionnaire survey is expensive to repeat.

The latter category (Homes and Solomon, 1951) reported that the human intelligibility level for a term correlates with its relative frequency of appearance in a balanced corpus of the written language. In other words, statistical term frequency can be an indicator for the intelligibility of the term. However, this statistical term frequency has the same disadvantages as the studies in the former category: valid only for the average person; difficult to cover technical terms; obsolete with time.

We therefore aim to develop a method that can solve the above problems.

3. Approaches

There are two important points in our proposed solution to the conventional problems. One is the construction of a term intelligibility database which includes not only subjective intelligibility scores but also the detailed attributes of each rater'. The other is the development of a method which can estimate any term intelligibility from an intelligibility survey of a limited number of technical terms. In this section, we discuss the above two approaches in turn.

3.1. Term Intelligibility Database with Raters' Attributes

Term intelligibility cannot be an absolute measure since it depends on the reader's knowledge. On that account, a term intelligibility database needs to offer not only subjective intelligibility scores but also each rater's detailed attributes (such as occupation, areas of interest and main source of information). Moreover, the database needs to compile information on several hundreds of people in order to offer various target reader models with regard to term intelligibility. However, for the conventional intelligibility database described in section 2., the number of raters is very few, and the information on raters' attributes is not sufficiently detailed.

3.2. Estimating Intelligibility from Limited Data

A subjective assessment to cover the enormous number of technical terms and to keep up with the expansion of technical terminology costs too much. Therefore, we need to develop a method which can estimate any term intelligibility from limited survey data.

In order to estimate this from limited survey data, we assume the following two points.

- Terms that have similar intelligibility share similar distributions.
- The term frequency weighted by the type of document, that is the term distribution, can be an indicator which reflects the target readership's term intelligibility.

The reason for the former is that terms that are unintelligible to many people occur more frequently in special documents than in general documents, and terms that are intelligible to a large number of people occur frequently in both special and general documents. The reason for the latter is that the term distribution in each document reflects the readership. For example, technical writing such as in scientific articles, includes more technical terminology than general text such as general newspaper articles.

We therefore propose a measure of intelligibility conditioned on reader clusters which is estimated by a regression analysis of the term frequency in different types of documents.

4. Sample Data and Evaluation Data

4.1. Survey Data Used for Regression Analysis

On the basis of the discussion in Section 3.1., we conducted a questionnaire on the subjective intelligibility rating of technical terminology. In the survey, we chose 46 technical terms from technical papers in the following fields: electricity, structural engineering, and environmental science. We categorized about 1000 respondents into three types laymen, engineers, and researchers, by the type of their main sources of information: general newspapers, trade newspapers and academic journals respectively. The intelligibility level of each term was rated on a five-point scale (1: quite intelligible, to 5: quite unintelligible).

We also asked the respondents (monitors for a marketing research firm) for characteristic details about themselves, such as their level of interest in new technologies, main sources of information on new technology, etc. The firm provided other details such as age, sex, place of residence, type of job, place of work etc. for analysis of the reader models.

4.2. Term Frequency Used for Regression Analysis

On the basis of the discussion in Section 3.2., we calculated the term distributions from term frequencies in every domain type on the Web, because it covers an enormous number of technical terms and is updated every day.

For information on the term frequencies in various types of domains, we used the number of documents that the Internet search engine "Google" retrieved using the term as a keyword. Google can specify the kind of website and search within websites by identifying the site's domain name ending. For example, domain names for Japanese universities/colleges end in "ac.jp," and domain names for general Japanese Internet Service Providers end in "ne.jp." If we want to search within Japanese academic/general websites, we can use Google to search within websites ending in "ac.jp" or "ne.jp." Therefore, we check the term frequencies of the 46 terms within "ne.jp" and "ac.jp" websites as examples of Japanese general and academic websites respectively. If we increase the types of website, the search time can increase even though the regression error is expected to reduce. Therefore we limited the types of website to two.

4.3. Survey Data Used for Evaluation

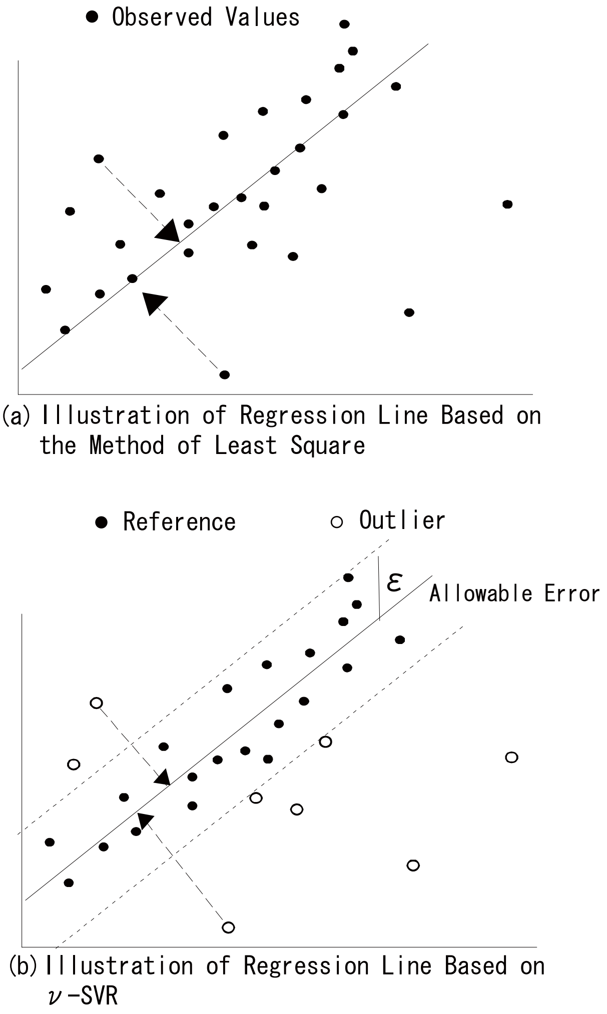
In order to evaluate our proposed model as described in Section 5. and 6., we chose 50 technical terms from technical papers in the fields of physics, electrical engineering, materials science, and meteorology. In this survey, the intelligibility level of each term was rated on a three-point scale to reduce the survey costs. The ratings on the three-point scale were translated to ratings on a five-point scale for the evaluation. We also asked the respondents for characteristic details about themselves as in the case of Section 4.1..

5. Nonlinear Regression Model

As the first pilot study, we obtained the following regression equation for estimating the layman's intelligibility level for a term based on the sum of the logarithms of the term frequencies; that is, the number of documents retrieved using the term from various types of websites:

$$y = f(\boldsymbol{x}) = \sum_{s \in S} w_s \log(x_s + 1) + b \qquad (1)$$

where y is the intelligibility level of the term, S is the set for that type of website, x_s is the number of documents retrieved using the term x from the websites s, and w_s is the weight for that type of website. We assume that a person's intelligibility level for a term is proportional to the logarithm of the number of times the person comes into contact with the term, referring to Weber-Fechner's law (Weber, 1834) (Fechner, 1860) "the magnitude of psychological sense is proportional to the logarithm of the magnitude of physical stimulus." Thus we use not x_s, but its logarithm $z_s = \log(x_s + 1)$ in Equation (1). Accordingly, this model is a linear expression of the form . $g(\boldsymbol{z}) = \sum_{s \in S} w_s z_s + b$. In order to obtain our regression model, we used the data for the intelligibility for the 46 technical terms (described in Section 4.) as y_i, and used the term frequencies within websites of type s as $x_{s,i}$.



(a) Illustration of Regression Line Based on the Method of Least Square

(b) Illustration of Regression Line Based on ν-SVR

Figure 1: Difference in Regression Procedure

In consideration of some outliers which differ considerably from the estimated value, we use ν-SVR (New Support Vector) (Scholkopf et al., 2000) to obtain the above regression equation from the sample data, instead of the method of least squares. In the method of least squares, the regression equation obtained depends largely on outliers because the method minimizes the total error (illustrated in Figure 1(a)). In ν-SVR, observed values lying within $g(\boldsymbol{z}) \pm \epsilon$ are regarded as accepted reference values; values outside the range are regarded as outliers (illustrated in Figure 1(b)). If the number of sample data$\{(\boldsymbol{z}_i, y_i)\}$ is N and the number of outliers is νN, ν-SVR determines the estimation equation $g(\boldsymbol{z})$ and ϵ which minimizes the sum of the mean average errors of outliers $\sum_i |y_i - g(\boldsymbol{z}_i)|/(\nu N)$, and $1/(2C)\|w\|^2$. The second term $1/(2C)\|w\|^2$ controls the stability of estimation against the fluctuation of input data. The numerical constant C determines the balance between robustness and error. In order to get the values for each regression coefficient, we set the number of outliers (νN) to 15 (30% of the sample data) in reference to the results of an exploratory study.

5.1. Evaluation

To evaluate the above model, we compared the estimated intelligibility level of the 50 technical terms with the average of the actual ratings from the survey result (described in the Section 4.3.) . Figure 2 is a scatter diagram of the estimated and the average actual intelligibility rating for each term. The correlation coefficient between the estimated rating and the average actual rating for the 50 technical terms is 0.66. This result shows that our model based on ν-SVR can estimate lay readers' intelligibility levels with an acceptable degree of reliability.

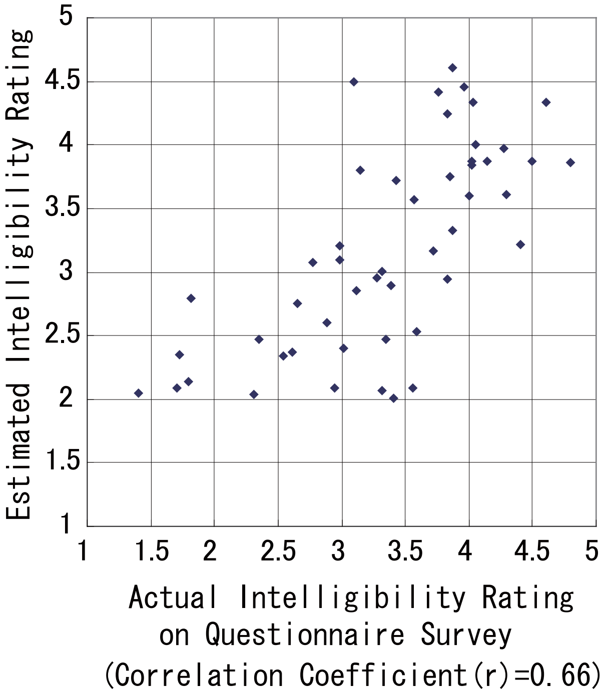


Figure 2: Estimated & Average Actual intelligibility Ratings for ν-SVR

Moreover, in a separate study, we developed a system that we call a "technical term checker" based on our model using ν-SVR. The checker alerts the user that the entered term might be too difficult for lay readers if the estimated rating of the term is more than 2.5. We introduced the checker to our title revision assistant system (Senda et al., 2004), which assists in revising a draft title in order to compose a title comprehensible to lay readers.

In order to test the effect of the title revision assistant system inluding the checker, we conducted an experiment which had 17 technical researchers revise their titles. After the revision, we conducted a questionnaire survey on the comprehensibility of the revised titles and terms to lay readers. As a result of the experiment, we confirmed that the checker enabled the researchers to use more comprehensible terms for lay readers than before. Moreover the result of the questionnaire asking for feedback from researchers showed that the checker was well received among them.

6. Linear Regression Model

As the second pilot study, we use multiple regression to obtain regression Equation 2 from the same sample data.

We define the term intelligibility y_r as the weighted sum of the logarithms of the term frequencies over all domain types:

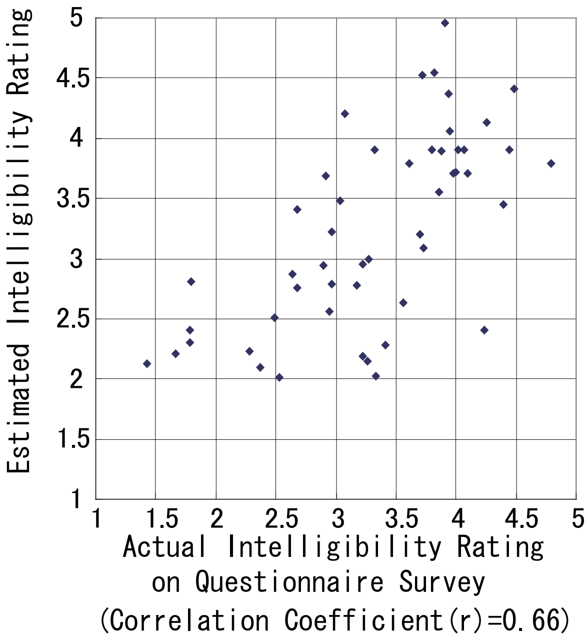


Figure 3: Estimated & Average Actual intelligibility Ratings for Multiple Regression

$$y_r = \sum_{s \in S} w_{s,r} \log(x_s + 1) + b + \epsilon \qquad (2)$$

where y_r is the term intelligibility level for a readership r, S is the set of domain types, x_s is the term frequency in domain s, $w_{s,r}$ is the weight of domain s for readership r, b is the intercept value, and ϵ is an error term. The term intelligibility level is proportional to the logarithm of the frequency, referring to Weber-Fechner's law (Weber, 1834) (Fechner, 1860).

Weights w_s can be obtained by a multiple linear regression analysis to fit the survey data described above using the term frequencies for every domain type for all the terms which are included in the survey data. Therefore, if we check the term frequencies for a specific term in every domain type, we can obtain its term intelligibility level. Additionally, if we also include the characteristics of the reader, we can obtain a more precise value for the term intelligibility level.

6.1. Evaluation

To evaluate the above model (based on multiple regression), we compared the estimated intelligibility level of the 50 technical terms with the average of the actual ratings from the survey result (described in the Section 4.3.) . Figure 2 is a scatter diagram of the estimated and the average actual intelligibility ratings for each term. The correlation coefficient between the estimated rating and the average actual rating for the 50 technical terms is 0.66. This result shows that the model based on multiple regression can also estimate lay readers' intelligibility levels with an acceptable degree of reliability.

The above two pilot studies show that these two regression models can estimate lay readers' intelligibility levels with a tolerable degree of reliability. In other words, our approach focusing on a reader model and the term statistics seems to be promising.

7. Conclusions and Future Work

This paper describes automatic terminology intelligibility estimation for readership-oriented technical writing. This paper assumes that term frequency weighted by types of documents can be an indicator of the term intelligibility for a certain readership. From this standpoint, we analyzed the relationship between the following: average intelligibility levels of 46 technical terms that were rated by about 120 laymen; numbers of documents that an Internet search engine retrieves using each term as a keyword from various types of websites (i.e. term frequencies). The results of the analysis show that term intelligibility for a target readership can be estimated by regression analysis of the term frequencies weighed by types of domain. As pilot studies, we developed two regression models estimating the technical term intelligibility for the target readership. One uses the machine learning method based on ν-SVR, and the other uses multiple regression.

In order to evaluate the models, we used the results of a survey on laymen's intelligibility levels for 50 new technical terms, and then compared the survey results with our estimated results. The results gave a correlation coefficient of 0.66 between the survey and the estimated results. The proposed model therefore is promising and has more potential with further analysis and development.

We are now analyzing more detailed information on reader models using the results of the questionnaire described in Section 4.1.. We plan to conduct experiments that will reflect the results of the analysis.

8. References

S. Amano and T. Kondo. 1998. Estimation of mental lexicon size with word familiarity database. In *Proc. of International Conference on Spoken Language Processing Vol.5*, pages 2119–2122, Sydney, Australia, December.

G. T. Fechner. 1860. Elemente der psychophysik 1 u. 2. (Breitkopf u. Hartel, Leipzig).

Davis H. Homes and Richard L. Solomon. 1951. Visual duration threshold as a function of word-probability. *Journal of Experimental Psychology*, 41:401–410.

Bernhard Scholkopf, Alex J. Smola, Robert C. Williamson, and Peter L. Bartlett. 2000. New support vector algorithms. *Neural Computation*, 12(5):1207–1245.

Yasuko Senda, Yasusi Sinohara, and Manabu Okumura. 2004. A support system for revising titles to stimulate the lay reader's interest in technical achievements. In *Proc. of 20th International Conference on Computational Linguistics*, pages 155–161, Geneva, Switzerland, August.

E. H. Weber. 1834. De pulsu, resorptione, audita et tactu. -annotationes anatomicae et physiologicae-. (Trs. by H.E. Ross, Academic Press, New York, 1978).