Method to Evaluate Difficulty of Technical Terms

Yuta Sudo[1(✉)], Toru Nakata[2], and Toshikazu Kato[1]

[1] Graduate School of Science and Engineering, Chuo University, Tokyo, Japan
all.rneh@chuo-u.ac.jp, kato@indsys.chuo-u.ac.jp
[2] National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan
toru-nakata@aist.go.jp

Abstract. We have developed an auto annotating system. To apply to the system, we conducted experiments about the method to evaluate difficulty of technical terms in documents by using data of Wikipedia. Based on a hypothesis that basic and easy terms appear frequently in Wikipedia, we surveyed relationship between subjective difficulty and appearance frequency in Wikipedia. As a result, we could classify technical terms into the easy term and the difficult term at the accuracy of 0.70.

Keywords: Word clustering · Automatic annotation · Information assistance

1 Introduction: Demand of Automatic Detecting of 'Difficult' Terms

Technical documents often contain technical terms without explanations, so non-expert readers may fail to understand the documents. To solve this problem, we developed an auto annotating system (Fig. 1).

The system should automatically detect 'difficult' technical terms and attach explanations on them. Evaluation of difficulty of terms is not so trivial.

This paper presents a method to score difficulty of technical terms by analyzing terminological structure of Wikipedia. The method evaluates difficulty of each term by observing appearance frequency of terms.

2 Related Work

Several researches on evaluation of easiness (or familiarity to readers) of terms have been conducted. Amano et al. [1] proposes an evaluation method that employs a catalog of familiarity of words. YAGO [2] should be mentioned as a famous example of ontology dataset. Such ontologies may be used to rank difficulty of terms, because we can guess that terms connected to difficult terms are difficult too.

Those methods, however, can work only on fixed vocabulary. We propose a method that can automatically generate the catalog of difficulty. This catalog will play effective role in the field of Document's Readability Assessment [3, 4].

S. Yamamoto (Ed.): HIMI 2016, Part II, LNCS 9735, pp. 72–80, 2016.
DOI: 10.1007/978-3-319-40397-7_8

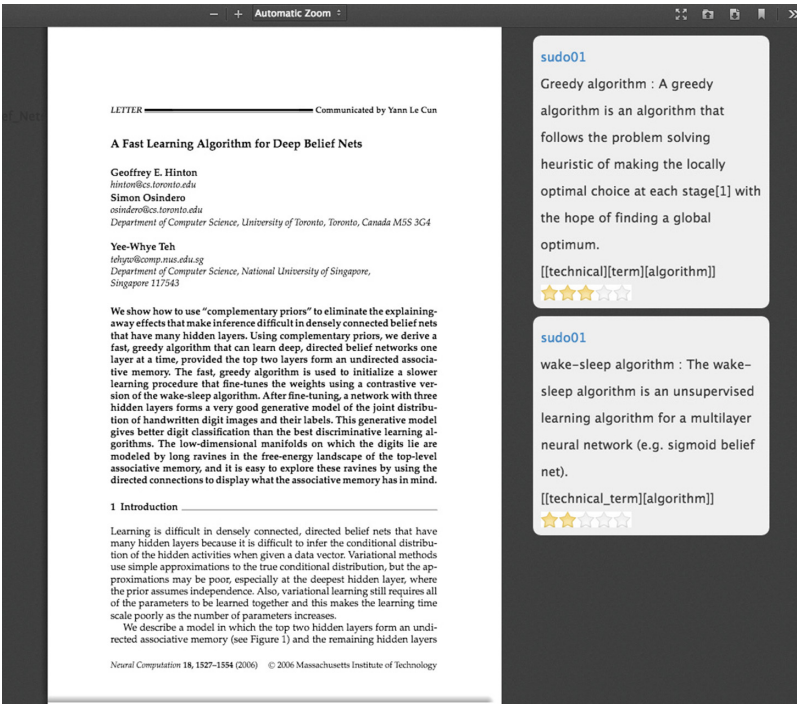

Fig. 1. Screenshot of our auto annotating system

3 Our Hypothesis on Characteristic of Term Difficulty

We define 'basic' and 'specific' as follows:

We call Term A is more basic than Term B when Term A is required to define Term B.

We define 'specific' as the antonym of 'basic'.

Our research starts with a hypothesis: "Basic terms tend to be referred very frequently in documents to define and to explain more specific terms."

4 Experiments

Impression of easiness of a term is subjective, while we defined the meaning of 'basic' objectively. We will investigate the correlation between easiness and basicness.

We performed the following experiments on the Japanese articles in the category of 'statistics' in Wikipedia.

4.1 Distribution of Term Appearance Frequency

We counted number of appearance of each index word within the 'statistics' category of Wikipedia. Figure 2 shows the distribution, and we find that easy terms appear more frequently than difficult terms in general. It corresponds to our hypothesis.

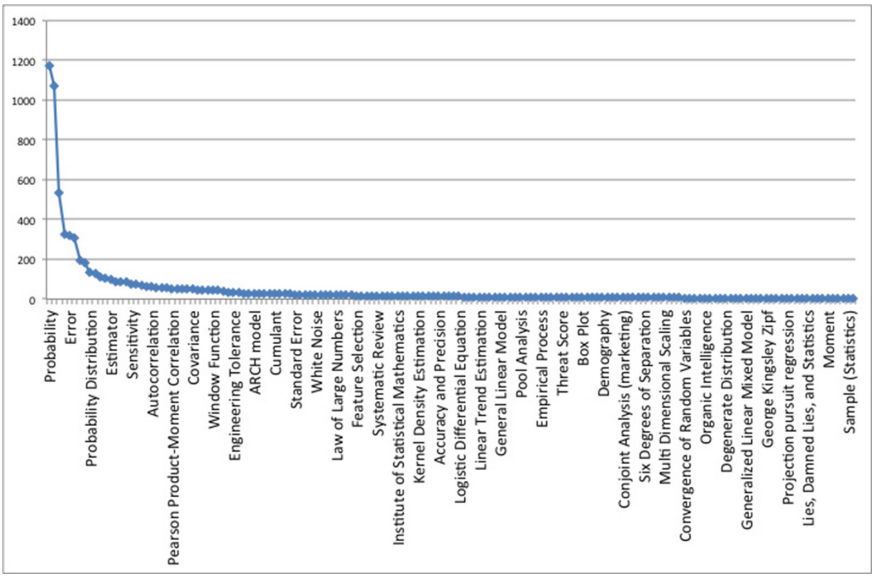

Fig. 2. Distribution of appearance frequency of index words

4.2 Relationship Between Frequency and Subjective Difficulty

We conducted a questionnaire research to measure subjective difficulty of the technical terms. We select 20 terms from the category (Table 1). We employ 11 people to rate difficulties of the terms. The scale of rating has 4-degree, and each subject answers by the number as following:

1. I have never heard the term.
2. I have heard the term, but I do not know the meaning of it.
3. I know the meaning of the term to some extent but not deeply to explain to the last detail.
4. I know the meaning of the term deeply.

The result shown in Fig. 3 indicates the correlation between appearance frequency and subjective difficulty. The correlation coefficient was 0.71. The fitting line for Fig. 3 was $y = 0.24x + 1.37$. It supports our hypothesis to some extent.

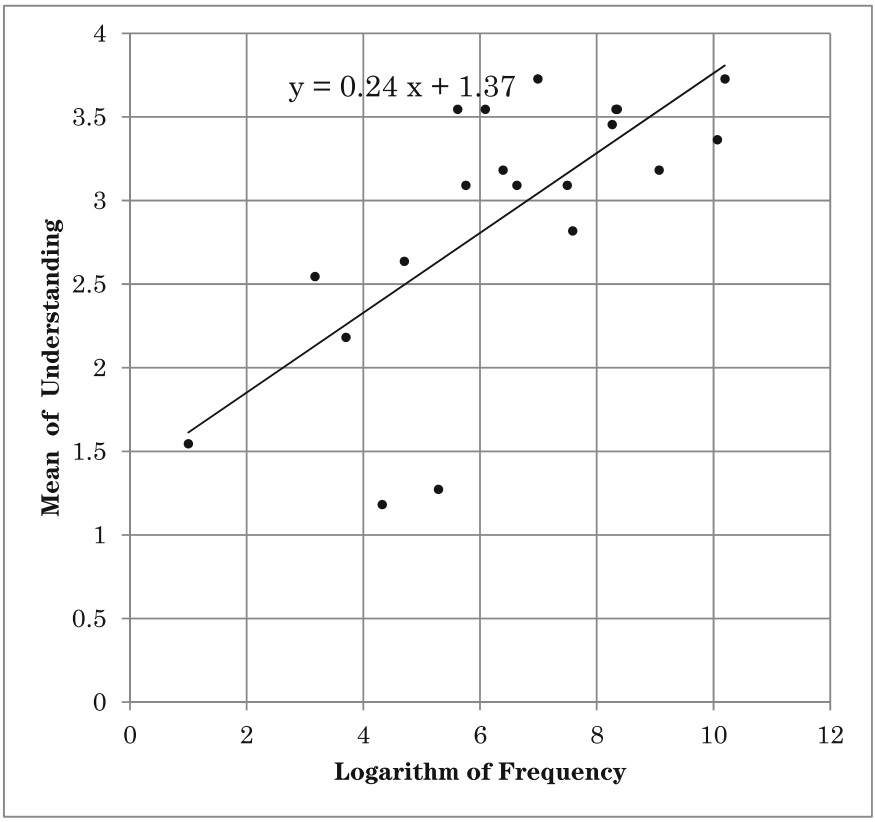

Fig. 3. Appearance frequency and averages of subjective score of understanding.

4.3 Agreement of Subjective Word Difficulty and Its Estimation Based on Word Frequency

In Sect. 4.2 we got the formula to estimate difficulty of words based on frequency of terms. We evaluate the accuracy of the estimation formula.

Considering the meaning of "3. I know the meaning of the term to some extent but not deeply to explain to the last detail" in questionnaire, we classified terms whose estimation of understanding is lower than three as the difficult terms that should be annotated. On the contrary, we classified terms whose estimation of understanding is equal or higher than three as easy terms that should not be annotated.

Table 2 is the confusion matrix between the estimated term difficulty and the subjective difficulty judged by the subjects. Fourteen words out of twenty are estimated correctly.

Table 1. Subjective degree of understanding and appearance frequency of terms

Words	Understanding	Frequency	Logarithm of frequency
Probability	3.73	1172	10.19
Statistic	3.36	1071	10.06
Statistics	3.18	536	9.07
Inference	3.55	326	8.35
Error	3.55	321	8.33
Random variable	3.45	307	8.26
Uncertainty	2.82	192	7.58
Deviation	3.09	180	7.49
Statistical population	3.73	127	6.99
Estimator	3.09	99	6.63
Probability density function	3.18	84	6.39
Standard score	3.55	68	6.09
Miracle	3.09	54	5.75
Covariance	3.55	49	5.61
Probability mass function	1.27	39	5.29
Family budget research	2.64	26	4.70
True negative rate	1.18	20	4.32
Kernel density estimation	2.18	13	3.70
Homoscedasticity	2.55	9	3.17
Directional statistics	1.55	2	1.00

Table 2. Result of automatic estimation of term difficulty based on term frequency compared to impression of the subjects

	Subjectively difficult	Subjectively easy
Estimated as difficult	6	5
Estimated as easy	1	8

5 Discussion

5.1 Good Amount of Annotation for Users

Considering the meaning of questionnaire, we chose three as the threshold of necessity of annotation in Sect. 4.3. Off course, we should consider about other value for the threshold. Changing the threshold value, we got the curve of classification accuracy (Fig. 4). The estimation based word frequency achieved 60 % accuracy in any threshold.

5.2 Referring/Referred Index and Subjective Difficulty

As a new candidate of difficulty indicator, we employ the number of referred times and referring times of each index word.

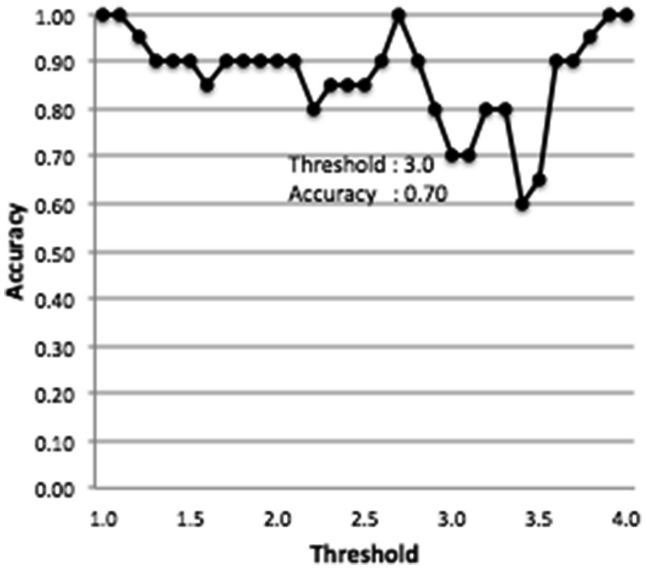

Fig. 4. Threshold of easy/difficult grade of words vs, accuracy of estimation based on word frequency

For example, we pick up the index word of "Least Squares Method (LSM)". This term refers (Table 3) and is referred (Table 4) by many other index words.

Table 3. Index words referred by term "LSM"

Title	Referred frequency	Frequency of appearance
Error	30	321
Least squares method	11	45
Residual	7	46
Probability theory	2	113
Independence (probability theory)	2	11
Presumption	2	326
Residual sum of squares	2	2
Probability	2	1172
Observation error	1	50
Statistic	1	1071
Covariance	1	49
Statistics	1	536
Non-linear least squares	1	6
Deviation	1	180
Maximum likelihood estimation	1	20
Degrees of freedom (statistics)	1	59

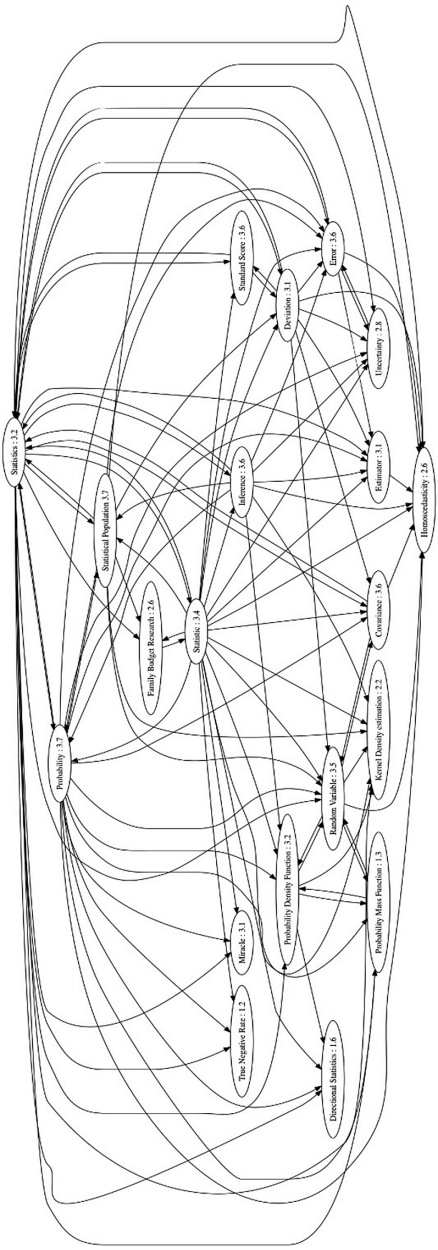

Fig. 5. Referring/referred relationship network. A term in upper stream is referred by the Wikipedia article of the term in lower. Numbers in each oval are subjective degree of understanding of term.

According to our definition of 'basic' and 'specific', index words in Table 3 are more basic than the term LSM, and the terms in Table 4 are more specific.

The referring/referred relationship formulates a network structure among terms. Figure 5 is an example of such network among the index words concerning 'statistics' in Wikipedia. We call the network "RR Network".

Table 4. Index words referring term "LSM"

Title	Referring frequency	Frequency of appearance
Non-linear least squares	18	6
Least squares method	11	45
Linear trend estimation	6	11
Error	2	321
Regression analysis	2	55
Coefficient of determination	2	8
Linear regression	2	28
Residual sum of squares	1	12
Regression analysis (multi variable)	1	22

Figure 5 suggests that the words being in upper stream are regarded as 'basic' by the definition are also subjectively easy. Using this RR Network, we can estimate difficulty more accurate.

We defined a value of "reference rank" of a term as the difference between referred frequency and referring frequency. The correlation coefficient between subjective difficulty and A/(A + B) was 0.56, where A is the number of referring and B is the number of being referred.

6 Conclusion

We conducted experiments that indicating relation between term frequency and subjective difficulty. From the result, we classified terms. The accuracy was 70 %. Although we used three as the threshold, there is room for more research about the annotation amount that people needs. As the method to improve accuracy, we are considering using RR-network.

Acknowledgement. This work was partially supported by JSPS KAKENHI grants (No. 25240043) and TISE Research Grant of Chuo University.

References

1. Amano, S., Kondo, T.: Estimation of mental lexicon size with word familiarity database. In: Proceedings of International Conference on Spoken Language Processing, vol. 5, pp. 2119–2122 (1998)
2. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: a core of semantic knowledge. In: Proceedings of the 16th International Conference on World Wide Web, pp. 697–706. ACM, May 2007
3. Jiang, Z., Sun, G., Gu, Q., Bai, T., Chen, D.: A graph-based readability assessment method using word coupling
4. Sato, S., Matsuyoshi, S., Kondoh, Y.: Automatic assessment of Japanese text readability based on a textbook corpus. In: LREC, May 2008