対象読者の専門度合に応じた用語の難易度の推定

千田恭子, 篠原靖志, 小杉素子, 長谷川尚子
(財) 電力中央研究所
senda, sinohara, kosugi, naoko-h@criepi.denken.or.jp

1　はじめに

ある専門用語が、どの程度専門的で難解か、もしくは意外に知られているかは、自分の専門分野に近いほど客観的に判断しづらいものである。その上、専門用語を使って行う説明の受け手の専門知識の度合 (レベル) がその時々で異なる場合、受け手ごとに分かりやすい用語を選択し使い分ける必要がある。

たとえば企業・行政等の大きな組織では、説明責任を果たす際、専門的な事柄でも、専門外の受け手にも分かりやすく説明する必要がある。ただし、その説明の受け手は、顧客、地域住民、環境団体、株主など、その時々で専門知識のレベルが異なることがある。またたとえば教育においては、初級者や中級者に向けて専門分野を説明する。その際、受け手の専門知識のレベルは段階に応じて異なる。それぞれの場面で説明の仕方を考える際、専門用語の難易度を、受け手の専門知識のレベルに応じて客観的に推定できると有用である。

そこで本論では、アンケートによる用語の難易度調査の結果と、ウェブ上の語の出現頻度との関係を分析し、受け手の専門知識のレベルに応じて用語の難易度を推定する手法について検討する。

2　従来の研究

用語の難易度の推定に関する従来研究は、主に 2 つに分けられる。一つは、多数の人に質問紙調査で用語の難易度を尋ねて 5-7 段階で主観的に評価させ、その平均値から難易度を明らかにする研究 [1] である。この研究においては、質問紙調査等で調べられる語の数が限られるため、未調査語が生じる問題がある。また、用語調査の結果は時間と共に陳腐化してしまう問題点もある。

もう一つの研究は、文書等での出現頻度数からその語の認知度 (親密度とも呼ばれる。これも一種の難易度と考える) を客観的に推定する研究 [2] である。この研究では、(新聞・雑誌・小説等の) 一般的な文書における常用語の出現頻度と、一般の人のその用語の認知度 (難易度) とを調べ、それらに相関がある事を指摘している。つまり、語の出現頻度はその読者層の認知度を反映している。従って、説明の受け手がよく目にする文書での用語の出現頻度を調べれば、受け手にとっての難易度の指標となる可能性がある。しかし、そのような文書の見極めが容易とは限らない。また仮に容易だとしても、その文書での出現頻度数の調査が、データの入手の困難さやコスト等の関係で難しい場合もある。

上記の先行研究を鑑みて、筆者らはこれまで、説明対象とする受け手が親しむ文書へのアクセスがしにくい場合でも、その受け手にとっての専門用語の難易度を推定でき、なおかつ未調査語や陳腐化の問題を回避可能な手法の確立を目指してきた。そして、専門用語の主観的な難易度と、ウェブ上の用語の出現頻度との関係を学習することで、一般の人にとっての専門用語の難易度を、ウェブ上の出現頻度から推定できる事を示した [3]。

しかし、既提案手法で導出する推定式では、用語の難易度を、専門知識のレベルが異なる受け手に応じて推定したい場合にも有効であるかは未検討である。そこで本研究では、専門知識のレベルが異なる受け手ごとの用語の難易度推定手法の確立を目指す。

3　知識レベルの異なる評価者ごとの評価値と、出現頻度の種類

本研究では、専門用語の主観的な難易度と出現頻度との関係を、専門知識のレベルの異なる評価者群ごとに学習することで、異なる評価者群に応じた用語の難易度を、出現頻度数から推定できる事を検討する。また、推定に適した出現頻度を調べる対象文書を検討する。

この節の以降では、この検討について「なぜ専門知識のレベルで学習を分けるか」「出現頻度を調べる対象文書の種類」の 2 点から順に説明する。

3.1　難易度と評価者の知識レベル

用語の難易度は絶対的なものではなく、読み手の知識レベルに応じて変動する。そのため、用語の難易度の主観的な調査を行う際には、難易度の評価値だけでなく、評価者の専門知識のレベルを示唆する属性情報 (主な情報源、職業など) も集めるべきである。そして、知識のレベルが異なる評価者の別に評価値を扱うべきである。そこで本研究では、知識レベル別に集めた評価者に、専門用語の難易度を評価してもらう調査を実施する。

3.2　出現頻度を調べる対象文書の種類

用語の認知度とその出現頻度との相関は高く、出現頻度は語の難易度を示唆する指標となる。ただし、分

野や知識レベルの異なる読み手の難易度の指標とするには、どんな種類の文書の出現頻度を参照すると良いかは明らかでない。本研究では、この点を検討する。

3.3 研究方法

本研究では、この節の上記で述べたことを検討し、難易度推定手法を確立するために、調査対象とする専門分野と、難易度を調べる専門用語をまず選定する。次に、出現頻度の種類を検討し数を調べる調査と、各用語の難易度を調べる調査とを行う。最後に、用語の出現頻度数と難易度との関係を、専門知識のレベルの異なる評価者ごとに分析することで、情報の受け手の知識レベルに応じて難易度を推定する式を導出する。

4 分野と用語の選定

調査対象分野としては、当所の報告書や論文等から用語のデータを集めやすい環境科学、土木建築、送配電、原子力の 4 分野を選定した。その次に、調査費用等の関係で、各分野ごとに 90 語を選定した。

まず、当所の報告書や論文データベース等の登録キーワード、並びにサイト上の用語集から用語を収集した。

その次に、難易度の点で、できるだけ多様な用語を集めた調査を行うために、2 節で言及した既提案手法を用いて、各語の難易度を 1 から 5 までの数値で推定し、難易度のレベル 1.5 未満、2.0 未満、2.5 未満、3.0 未満の語を各 15 語、3.5 未満、4.0 未満を各 12 語、4 以上の語を 6 語、ランダムに選定した。レベル別の選択語数が示すように、難易度 3.0 以上の用語数は段階的に少なくしている。難解な用語が多い場合、アンケート調査において、専門家でない回答者に負担がかかり、回答に影響が出る事を防ぐためである。

5 出現頻度に関する調査

この節では、どのような文書での出現頻度を参照するのが良いかの検討と、出現頻度数の計測方法について説明する。

5.1 出現頻度を調べる文書の種類

先行研究では、常用語について、(新聞・雑誌・小説等の) 一般的な文書での頻度と認知度との相関を指摘している。ある専門分野の専門外の人全般がよく目にするのは、上記のような一般的な文書であると考えられるため、本研究でも一般の文書を、参照候補の一つとする。また次節で後述する理由により、専門的な文書での出現頻度も、他の情報と組合せて利用する。そこで、専門的な文書での出現頻度の情報も、参照候補の一つとして比較に用いる。

5.2 専門的な文書と一般的な文書での頻度分布

前節では、単一種類の文書の候補を挙げたが、ここでは複数種類の文書の組合わせについて述べる。専門的な文書と、一般的な文書とのそれぞれにおける出現頻度や、それらの相対的な関係は、その語の難易度を示唆する指標となる可能性がある。たとえば、難解な専門用語であればあるほど、一般的な文書に比べて専門的な文書での出現頻度数が大きい可能性が高い。また、平易な用語であればあるほど、一般的な文書と、専門的な文書での頻度の差異が小さい、もしくは一般的な文書における出現頻度数の方が大きい可能性が高い。つまり、専門度の異なる文書での出現頻度や、それらの相対的な関係は、その語の難易度を示唆する指標となる可能性がある。そのため、専門的な文書と、一般的な文書とのそれぞれにおける出現頻度数の関係も参照候補とする。

2 節で言及した既提案手法では、この観点に基づいて、専門的な文書と、一般的な文書とのそれぞれにおける出現頻度数を参照している。この手法による推定式は、一般の人の難易度については、一定の精度で推定できていた。

5.3 出現頻度数の計測

本研究では、用語の出現頻度を簡便に計測するため、ウェブ上の検索エンジンによる用語の検索件数を用いる。ウェブ上では多数の文書が公開されているが、そこでの検索エンジンによる用語の検索件数は、ウェブ上で計測できる一種の出現頻度数とみなせる。また、検索エンジンによる検索件数 (頻度数) は、学術機関 (ac.jp)、プロバイダー (ne.jp) といったドメイン名の種別ごとに算出できる。そこで、専門的な文書での出現頻度として、ac.jp サイト での検索件数を利用する。また、一般的な文書での出現頻度として ne.jp サイト、

上記で挙げたサイト上での検索件数を用語の頻度数とみなして用いる事には以下の利点がある。

- 特定の分野に依存しないため、対象とする分野が変わっても、共通の参照先として利用できる。
- 随時更新されるため、陳腐化の問題が避けられる。
- 専門用語に多い複合語や新語も随時記載されるはずである。仮にどこにも掲載されていない語があった場合は非常に専門性が高く、特殊で難解な用語であるとみなせる。従って未調査語は事実上ないと言える。

6 アンケートによる難易度調査

6.1 目的と調査方法

このアンケート調査の目的は、専門用語について、専門知識のレベルの異なる評価者群の別に評価した難易度を得ることである。そのために、まず専門知識のレベルを調べる予備調査を行い、その結果で選定した回答者に、用語の難易度を 5 段階で評価してもらう本調査を行った。

なおこの調査は、調査会社からの回答依頼の電子メールに応じたモニターに対して、Web ページ上の予備調査、本調査の質問紙に回答してもらい、実施・回収を行っている。

以下で、予備調査、本調査の概要を述べる。

6.2 予備調査：回答者の選定

4 節で選定した分野ごとに、専門知識のレベルの異なる 2 タイプの回答者を選定するための予備調査を行った。予備調査では、「○○についての情報を見聞きすることはありますか?あれば、どこで見聞きされるか、以下からお選びください」という質問を提示し、下記の選択肢を提示した (質問、選択肢とも、○○の部分は分野名を示す表現が埋め込まれる)。

a 特に見聞きすることはない

b 科学雑誌 (たとえば日経サイエンス、ニュートン、科学 (岩波) など)

c ○○に関する業界新聞や雑誌

d ○○に関する学会誌または論文誌

そして、選択肢の a を選んだ人から本調査の対象者をランダムに約 300 名選出し、関心がないため知識レベルが低めである層として、無関心層と名づけた。また、b または c を選択し、なおかつ d を選択しなかった人からランダムに約 300 名選出し、関心も知識もそこそこあるが、専門家ほどの知識はない層として、関心層と名づけた。本調査では、各分野ごとに、無関心層、関心層の二タイプの回答者を選定した。

ただし、回答者一人に 90 語の用語を割り当てては、一人当たりの評価語数が多く負担がかかりすぎ、回答に影響が出る恐れがある。そこで各回答者には、一人一分野につき 30 ／ 90 語を割り当てた。そのため、90 語をランダムに 30 語× 3 の三つの用語カテゴリに分割した。そして、各カテゴリを異なる 100 人に担当させるため、各タイプの回答者 300 人をランダムに 100 名ずつ各用語カテゴリに割り当てた。従って、100 名× 3 用語カテゴリ× 2 タイプの回答者× 4 分野で、計約 2400 名の回答者を選定した。

6.3 本調査：難易度の 5 段階評価

本調査では、前節で述べた通り 30 語の用語を提示し、各語について、その難易度を 5 段階の選択肢 (1. 非常によく分かる、2. だいたい分かる、3. なんとなく分かる、4. あまりよく分からない、5. 全然分からない) から選択してもらった。但し、用語の提示順序は、回答者ごとにランダムにした。

7 難易度推定モデル

これまでの研究 [3] を踏まえてサイト種別毎の検索文書数に基づき、用語の難易度を推定する予測式として、以下の形式の回帰式を、各分野の無関心層／関心層の別に求めた。

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

ただし、y は用語の難易度評価値を、S はサイト種別の集合を、x_s はサイト種別 s での用語 x の検索件数を、w_s は各サイト種別の重みを指す。予測式では「感覚は刺激の対数に比例する」という Weber と Fechner の法則 [4][5] にヒントを得て、難易度は、人が用語に接する頻度の対数に比例すると仮定し、検索文書数 x_s 自身ではなく、対数を取った $z_s = \log(x_s + 1)$ を使用している。 予測式は、$\boldsymbol{z} = (z_s)$ の一次式 $g(\boldsymbol{z}) = \boldsymbol{w}'\boldsymbol{z} + b$ となっている。

予測式の各パラメータは、(重) 回帰分析で決定する。回帰用データとしては、各分野の無関心層／関心層の別に評価した専門用語 90 語の難易度評価値 y_i と、それらの用語のサイト種別 s ごとの検索文書数 $x_{s,i}$ とを用いた。

専門的なサイトと、一般的なサイトとの両方での検索件数を使用する場合、サイト種別 (s) を増やせば回帰誤差は小さくなるが、サイト種別毎の検索が必要となり検索に要する時間も増える。このため、使用するサイト種別は最高 2 種類に限定する。つまり、5.3 節で述べたように、ac.jp など専門的なサイト、ne.jp など一般的なサイトの両方か、それらのサイトのうち一つを単独で利用する。

7.1 評価結果

まず、難易度推定モデルに出現頻度を変数として取り込む際、どのような文書の出現頻度数を推定式に組み込めばよいかについて議論する。そのために、ac.jp サイトでの出現頻度数、ne.jp サイトでの出現頻度数、両サイトそれぞれの出現頻度数を用いた場合とで、推定モデルを比較した。以降では、各モデルを順に AC、NE、AC&NE と呼んで参照する。どのモデルが最も良いか調べる基準として、各モデルの自由度調整済決定

係数、AIC、Mallow's CP の値を算出し比較した。その結果、4 分野× 2 タイプの回答者 (計 8 つのケース) のうち 7 つのケースで、NE が最も良く、AC が最も悪かった。AC&NE は、その 7 つのケースにおいて NE よりやや劣っていた。残りの 1 つのケースでのみ、AC&NE が、NE よりも良い結果を示していた。従って、多くの場合は、NE が最も良い。

次に、導出したモデルが、関心や知識のレベルが異なる層それぞれの難易度を予測できていたかについて述べる。図 1 に、送配電分野の無関心層・関心層の難易度について、実測値・予測値の関係を散布図に示す。図 1 の点 (実測値) の散布具合より、同じ 90 語について、二つの層は全く異なる評価値を答えていることが分かる。しかし、どちらの回答者層の散布点も $X = Y$ の線上付近に分布している。他分野の散布図においても、図 1 と同様に、二つの評価者層の散布具合は異なる分布を見せるが、$X = Y$ の線上付近に分布していた。このモデルで求めた推定式が、異なる反応を示す二つの層の難易度をどちらも推定できる事を示している。また、各分野の無関心／関心層について、予測値と実測値との相関係数を算出したところ、0.63 から 0.78 の値をとっていた。従って、このモデルによって求めた推定式は、異なる分野の異なる回答者層それぞれの難易度を一定の精度で予測できている事が分かる。

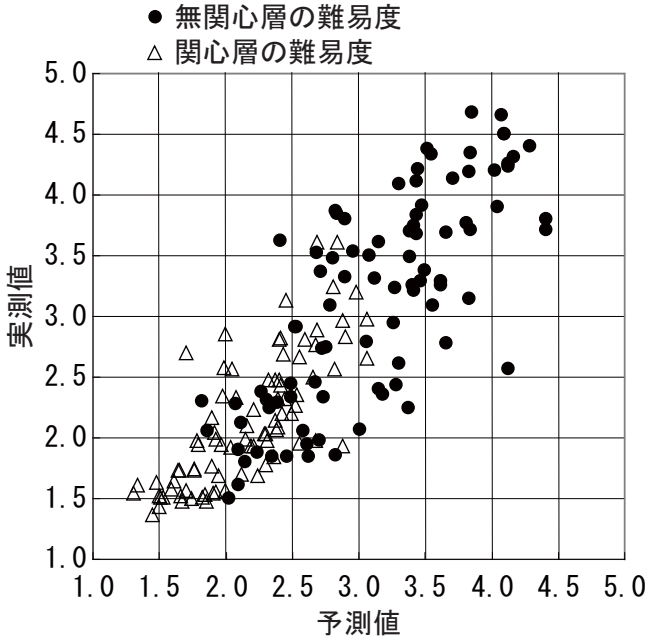

図 1: 無関心層／関心層の難易度の予測値と実測値 (送配電分野)

8 まとめ

本研究では、用語の難易度と出現頻度との関係を、専門知識の異なる評価者群ごとに分析することで、各評価者群の知識レベルに応じて用語の難易度を推定することができるか分析した。分析の結果、関心や専門知識のレベルが異なるために、同じ用語について異なる反応を示す評価者群のそれぞれについて、その難易度評価値を本提案手法は一定の精度で推定できた。

また、難易度推定モデルに変数として取り込む出現頻度について、専門的な文書 (ac.jp サイト) での出現頻度、一般的な文書 (ne.jp) での出現頻度、両文書それぞれにおける出現頻度のどれが良いかを分析した。それぞれの出現頻度を用いた推定モデルについて、自由度調整済決定係数、AIC、Mallow's CP の値を比較したところ、一般的な文書 (ne.jp) での出現頻度を参照するモデルがもっとも良いという事が分かった。

今後は、出現頻度以外にどんな情報を変数として組み込めば、推定式の精度をより高められるかについて検討する予定である。特に語構成等の情報の組み込みを検討している。

参考文献

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

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

[3] Yasuko Senda, Yasusi Sinohara, and Manabu. Okumura. Automatic terminology intelligibility estimation for readership-oriented technical writing. In *Proc. of the 5th International Conference on Language Resources and Evaluation*, pp. 1506–1509, Genoa, Italy, May 2006.

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

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