和英短文を対象とした著者専門性推定への応用

滝川 真弘† 山名 早人‡

†早稲田大学大学院基幹理工学研究科 〒169-8555 東京都新宿区大久保 3-4-1
‡早稲田大学理工学術院 〒169-8555 東京都新宿区大久保 3-4-1
E-mail: {temy0501, yamana}@yama.info.waseda.ac.jp

あらまし 本研究の目標は，特定分野に対する著者の専門性を如何に短い文章から判定するかにある.短い文章の例としては，質問投稿サイトの回答などがある.こうした短い文章が内包する特徴量は少ないため，既存研究では特徴量を増やすために，当該著者により記述された複数の文章や他の属性を用いて当該著者の専門性を推定している.しかし，当該著者に対して常に複数の文書や他の属性が用意できるとは限らない.この問題を解決するため，本論文では，著者の専門性を短い文章から推定することを目的とし，出現する単語に専門毎に適切な重みを付与する手法として「CrRv」を提案する．CrRv は「知名度の低い専門用語ほどその用語を用いる著者の専門性が高い」という仮定のもと当該用語に重みを付与する．そして，判定対象となる文章に CrRv の高い用語が含まれているほど，著者の専門性が高いと判断する．評価実験においては，従来から用いている Yahoo!知恵袋のデータ（対象特定分野は「医療」と「プログラミング」）と，WikiAnswers のデータ（特定分野は「プログラミング」）に対して，回答者の専門性の推定を行った.既存手法である tf-rf，tf-PNF2，tf-idfec_b との比較実験を行い Precision@10 により評価した．結果として，既存手法の中で最も性能の良い手法と比較して，日本語の場合で 0.2，英語の場合で 0.3，その絶対値を向上させることができた.

キーワード 単語, 重要度,

1. はじめに

本研究の目標は，十分な学習データを用意できない状態で，特定分野に対する著者の専門性を如何に短い文章から判定するかにある．短い文章とは，Twitter 等 SNS への投稿や，EC サイトでのレビュー，質問投稿サイトの回答などが挙げられ，いずれの場合も「特定分野における著者の専門性」は，投稿の信頼性などを判定する上で重要な要素となる．しかし，短い文章は一般的な文書とは異なり文量が少なく,情報量も少ない.具体的には出現する単語の種類や単語数が少なくなる.そのため，機械学習等の手法で精度を出すことが難しい[1].

こうした問題に対し，既存手法の中では機械学習を使わず，他の情報を用いて情報量を補い，推定する手法とっている．例えば，質問投稿サイトにおける専門性推定行なっている既存手法では，ユーザのつながりや貢献度[2][3]を用いるものや,あるユーザの複数の投稿をまとめて 1 つの文書として用いるもの[4][5]がある．しかしこれらの手法を適用するには，ユーザ自身の多くの情報が必要となる．したがって，新規ユーザやあまり活動していないユーザに対しては適用することができない．一方，文書の情報のみを用いて，機械学習を適用させる研究も存在する．Yang ら[6]は 2016 年に，ある 1 文書を深層学習を用いて分類する手法を提案している．これらの手法を応用することで専門性推定を行うことも考えられる．しかし，Yang らの結果によれば平均 91word からなる Amazon のレビューを 5 分類するタスク（本論文が対象とする専門性の分類は対象としていない）においての分類精度は 62.9%にとどまっている．また，機械学習を用いるには十分な学習データが必要となり，Yang らは学習のために 24 万から 240 万のデータを使用しており，新規サービスなど，データが十分にない状態での適用は困難である．

筆者らは，こうした文章以外の情報（ユーザ属性等）が十分にない場合にも有効に機能する手法として，1 つの短い文章のみから専門性を推定する手法に取り組んできた[7]．[7]においては，著者の専門性を短い文章から推定するため，単語の重み付与手法として CrRv を提案した．日本語の質問投稿サイトである Yahoo 知恵袋[1]において回答の著者の専門性推定実験を行い，有用性を示した．しかし，対象とした言語が日本語のみのため，他言語においての有用性を示せていない．そこで本稿では英語の短文に対して同様な評価実験をおこない，他言語に対して有用性を示す.

提案手法は，「適切な重みは対象とする分野毎に異なる」ことを前提とし,「特定分野を対象とした単語重要度」を計算する．提案手法は，特定分野における単語重要度を「一般人が使わない単語であり，かつ特定分野で用いられる単語の内，当該分野での出現頻度が低い方

1 https://chiebukuro.yahoo.co.jp/

がより重要度が高い」という仮説を前提に各単語に当該分野に対する単語重要度を付与する．具体的には，予め専門辞書が与えられている時，当該専門辞書内の単語を対象に重要度を付与する．重要度付与にあたっては，当該分野と当該分野以外のコーパスを用い，「当該分野以外のコーパスにはほとんど出現せず，かつ当該分野コーパスにおいても出現頻度の低い単語」に高い重要度を付与する．

以下，2 節にて関連研究，3 節にて提案手法．4 節にて実験に使用するデータセット，5 節にて評価方法，6 節にて実験結果を示し，7 節にて本稿をまとめる．

2. 関連研究

出現頻度と分野(カテゴリ)の観点から，単語の重要度を計算する手法について紹介する．

2.1 単語重要性を測る手法

文章中に表れる単語の重要性を測る手法としては，TF-IDF[8]が有名である．

TF-IDF[8]は，文書に索引を付ける際の重み付けを目的として考案された．TF-IDF は，ある文書集合中に存在する 1 つの文書における特徴的な単語を表現するために用いられるものであり，ある文書集合が与えられた際に，個々の文書を区別することのできる単語に高い重みを与える．具体的には，単語 t の文書 d に対する重要度 w(t, d)は，式(2.1.1)により計算する．TF(Term Frequency)は単語出現頻度であり，式(2.1.1)の tf(t, d)は，単語 t の文書 d 内での出現頻度を示す．DF(Document Frequency)は，単語が出現する文書頻度である．DF の逆数の値が IDF(Inverse Document Frequency)であり，この値が大きいと特定の文書のみに出現する傾向が高いことを示す．idf (t)は，式(2.1.2)により計算する．

$$w(t,d) = tf(t,d) * idf(t) \quad (2.1.1)$$

$$idf(t) = log\left(\frac{|D|}{df(t)}\right) \quad (2.1.2)$$

ここで，tf(t, d)は文書 d(∈D)中の単語 t の出現回数，|D|は全文書数，df(t)は単語 t が現れる文書 d の数である．

TF-IDF は，文章の検索インデックスなどに使用することを目的としている．すなわち，文書群に対する 1 つの文書内に存在する各単語の重要度を算出することにより，対象とする文書の特徴語を抽出している．このため，ある分野おける単語重要度算出のために直接用いることはできない．特定分野での重要度算出のためには，特定分野に属する文章集合を用意した上で TF-IDF を求めなければならない．しかし，特定分野に属する文書集合は，特定分野に関連しない単語を含んでいることから，特定分野に属する単語以外の単語にも大きな重みが与えられる可能性がある．

2.2 カテゴリと単語の関係から重要度を計算する手法

特定分野（カテゴリ）が付与された文書集合について，カテゴリに対する単語の出現頻度の偏りから重要度を計算する従来手法として，tf*rf[9]，tf*PNF^2[10]，tf*idfec-b[11]の 4 手法を紹介する．なお，以下の説明ではカテゴリ C に属する文書集合 Dp と属さない文書集合 Dn が用意されているものとする．さらに，単語 t に対して Dp のうち t が出現する文書数を a， Dp のうち t が出現しない文書数を b， Dn のうち t が出現する文書数を c， Dn のうち t が出現しない文書数を d，全文書数を N とする．

2009 年に Lan ら[9]は，ある文書がカテゴリ C に属するか否かを推定することを目的として，tf*rf と呼ばれる単語重要度計算手法を提案した．同手法は，単語 t の文書内での単語出現頻度 tf に加え，単語 t の出現か，〵あるカテゴリに属する文書集合と当該カテゴリに属さない文書集合でどれだけ異なるかを示す rf を用いる．具体的には，単語 t についての rf 値である rf(t)は，式(2.2.2)で表される．

$$rf(t) = \log\left(2 + \frac{a}{\max(1,c)}\right) \quad (2.2.2)$$

なお，tf(t, d)は文書 d 中の単語 t の出現頻度であり，tf*idf の tf と同値である．tf*rf は，tf(t, d)と rf(t)の積により求める．

一方， 2015 年に Behzad ら[10]は，tf*PNF^2 を提案した．Behzad らの目的も，ある文書がカテゴリ C に属するか否かを推定することである．Behzad らは従来の文書分類のための単語重要度計算方法は，文書集合 Dp，Dn の文書数に偏りがあると安定した精度が出ないことを指摘した．そこで a，b，c，d をそのまま用いるのではなく，Dp，Dn 内それぞれにおいて単語 t が出現する確率を求め計算を行う tf*PNF^2 を提案した．PNF^2 の式(2.2.3)に示す．なお，tf は，単語 t の文書内での単語出現頻度である．

$$PNF^2(t) = \frac{P(t_i \mid C) - P(t_i \mid \bar{C})}{P(t_i \mid C) + P(t_i \mid \bar{C})} \quad (2.2.3)$$

$$P(t_i \mid C) = \frac{a}{a+b} \quad (2.2.4)$$

$$P(t_i \mid \bar{C}) = \frac{c}{c+d} \quad (2.2.5)$$

Giacomo ら[11]は 2015 年に tf*idfec-b(t)を提案した．Giacomo らの目的も，ある文書がカテゴリ C に属するか否かを推定することである．Giacomo らは，カテゴリ分類において重要な要素は「ある単語 t が如何に該当カテゴリ以外で出現しないか」であると考えた．該当

†1 早稲田大学
Waseda University

カテゴリ以外での非出現割合に加えて該当カテゴリにおける文書頻度 a を組み合わせた tf*idfec-b を提案した．idfec-b を式(2.2.6)に示す．なお，tf は，単語 t の文書内での単語出現頻度である．

$$idfec\text{-}b(t) = \log\left(2 + \frac{a+c+d}{\max(1,c)}\right) \quad (2.2.6)$$

3. 提案手法[7]

我々は以前特定分野にどれだけ精通しているかを判断することを目的とした単語重要度計算手法を提案した[7]．ただし，前提条件として，特定分野に属する単語群（専門辞書）が事前に与えられているものとし，重要度（専門度）に応じて単語に重みを付与する．

提案手法のアイデアは，専門辞書には一般人も使用する単語（例えばプログラミングの場合、「java」）が含まれているのが一般的であり，専門辞書に含まれる単語の中でも一般人があまり用いない単語に高い重要度を付与することにある．つまり，特定分野にどれだけ精通しているかを判断するために，該当分野に精通していないと知り得ない単語に高い重要度を付与する．

上記を実現するために，特定分野のコーパス Dp と一般分野のコーパス Dn を使用する．そして、専門辞書に含まれる単語の内，Dn にはほとんど出現せず，かつ Dp 内でも出現頻度が低い単語ほど重要であるという仮説のもと，CrRv(Category relevance Rarity value)を提案する．以下，詳細を述べる．

3.1 CrRv

提案する CrRv を式(3.1.1)に示す．

$$\mathrm{CrRv(t)} = Cr(t) * IH(t) * TFMAX(t) \quad (3.1.1)$$

$$Cr(t) = \frac{DF_P(t)/|Dp|}{DF_P(t)/|Dp| + \alpha * DF_N(t)/|Dn|} \quad (3.1.2)$$

$$\mathrm{IH(t)} = \log\left(\frac{\max_{t' \in T} H(t')}{H(t)}\right) \quad (3.1.3)$$

$$H(t) = -\sum_{d \in D} P(t,d)\log P(t,d) \quad (3.1.4)$$

$$P(t,d) = \frac{tf(t,d)}{\sum_{d'}^{D} tf(t,d')} \quad (3.1.5)$$

$$TFMAX(t) = \max_{dp \in DP} tf(t,dp) - \max_{dn \in DN} tf(t,dn) * \beta \quad (3.1.6)$$

$$\alpha = \frac{\sum_{t'}^{T} DF_P(t')/|Dp|}{\sum_{t'}^{T} DF_N(t')/|Dn|} \quad (3.1.7)$$

$$\beta = \frac{\sum_{dp}^{DP}\sum_{t'}^{T} tf(t',dp)/|Dp|}{\sum_{dn}^{DN}\sum_{t'}^{T} tf(t',dn)/|Dn|} \quad (3.1.8)$$

上式において，対象とする単語を t，特定分野の文書集合を Dp，一般分野の文書集合を Dn，全文書集合を D（=Dp+Dn）で表す．全単語集合を T，Dp の文書の数を|Dp|，文書 d 中に出現する単語 t の数を tf(t，d)，単語 t の Dp における文書出現頻度を $DF_P(t)$，単語 t の Dn における文書出現頻度を $DF_N(t)$としている．また，α，β は単語 t がコーパス Dn に出現した際に重要度を下げる割合を調整するパラメータである．

式(3.1.1)において，Cr(t)は単語 t の当該カテゴリへの出現頻度の偏り具合を示し，当該カテゴリへの片寄りが強い単語に大きな重要度を付与する．IH(t)は単語 t が文書集合 D 中の各文書に異なる頻度で出現するほど大きくなる値であり，単語 t の文書集合 D 内での特異性を表す．すなわち，特異な単語ほど高い重要度を与える．TFMAX(t)は，IH(t)によってノイズ的な単語が大きな重要度を持つことを避けるための項である．以下，各々の項について詳細に説明する．

Cr(t)は、単語 t を持つ文書が特定分野コーパス Dp に属する文書群へどの程度偏っているかを示しており，Dp に偏っているほど大きな重要度を与える．ただし，|Dp|と|Dn|は同一ではないため正規化している．α は $DF_N(t)$の影響を調整するパラメータであり，設定方法については後述する．

IH(t)は，単語 t の全文書集合 D に対するエントロピーの逆数（単語 t ∈T の最大エントロピーで正規化している）であり，「文書集合 D 内の特定の文書に集中して出現するほど大きく」なる．すなわち少数の文書にしか出現しない単語に大きな重要度を与えている．このように，IH(t)を用いることで特異性のある単語に大きな重みを与えることができる．

TFMAX(t)は，ノイズとなる単語の重みを小さくするための項である．IH(t)により文書集合 D 中で特異性のある単語に高い重みを付与することが可能となるが，一方で偶然出現するノイズ的な単語（少数の文書のみに出現する単語）の重要度が高くなってしまう．そこでノイズとなる単語は「1 文書内の出現頻度が低い」ことに着目し，1 文書内での出現頻度が高い単語の重要度を上げることで相対的に出現頻度の低い単語の重要度を下げる．具体的には，単語 t の Dp 内での tf 値の最大値 $\max_{dp \in DP} tf(t,dp)$を用いる．一方，Dn 内で tf 値が高い単語は重要度を下げるべきであり，最終的に $\max_{dp \in DP} tf(t,dp)$ から $\max_{dn \in DN} tf(t,dn)$ を減じることで $TFMAX(t)$を計算し，重要度計算の一つのパラメータとした．ただし，$\max_{dn \in DN} tf(t,dn)$の影響を調整するため，式(3.1.6)に示す通りパラメータ β を付加している．

次にパラメータ α と β の求め方について示す．なお，これらのパラメータは，データセット Dp，Dn に依存

†1 早稲田大学
Waseda University

する値である．これは，Dp，Dn の何れの文章集合に含まれる文書についても，各々の集合に含まれるべき文書である確率は高いものの，必ずしも正しいとは限らないこととを考慮するために付加している．本研究では，α と β をいくつかの計算方法によりで検証し，その中で最もよい性能を出した計算方法を採用した．具体的な計算式を式(3.1.7)，(3.1.8)にて示す．

最終的に採用したαは，一般分野コーパス Dn 内の文書に比較して，特定分野のコーパス Dp 内の多くの文書が，単語 t を持つほど大きくなる．すなわち，式(3.1.2)から分かるように Dp 内の多くの文書が t を内包する場合に Cr(t)の重要度を下げている．一方，β は，Dp 内での単語 t の出現頻度が Dn 内での単語 t の出現頻度より大きいほど大きくなる．すなわち，式(3.1.6)から分かるように，Dp 内での単語 t の出現頻度が大きいほど TFMAX(t)を大きくし重要度を上げている．

3.2 CrRv を用いた著者専門性計算の流れ

本項では提案手法 CrRv(t)を使用するために必要なデータセットおよび CrRv(t)を用いた文章からの著者専門性計算の流れについて説明する．

使用するにあたって必要な入力(データセット)は前述の通り，特定分野コーパスと一般分野コーパス，それから計算対象となる関連用語の集合である．また，出力されるものは入力した専門用語それぞれに対して重要度を付与した辞書である．

続いて CrRv(t)を用いた文章からの著者専門性計算の流れについて説明する．CrRv(t)は単語 t に重要度を付与する手法であり，そのまま文章 x に対して文章 x の著者の専門性を付与することができない．そこで文章 x に対し，CrRv(t)を用いて後述する専門性スコア(x)を求め，著者専門性の計算を行う．

専門性スコア (x) について説明する．ある文章 x の専門性スコアを AnswerScore(x)とする．また，使用する専門辞書に含まれる単語集合を T とし，単語t_j($t_j \in T$, $1 \le j \le |T|$)が回答 x の中で出現したら 1，出現しなかったら 0 を出力する関数を exist (x, t_j)とする．単語t_j の重みは W(t_j)とする．単語の出現回数から生成した |T| 次元のベクトルを AnswerVec(x)=[exist(x,t1), exist(x,t_2) … exist(x,t_j) , … exist(x,$t_{|T|}$)], |T|次元の単語重要度ベクトルを WeightVec=[W(t_1), W(t_2), …W(tj), …W($t_{|T|}$)]とした時，AnswerScore(x)を式(3.2.1)に示す．

$$AnswerScore(x) = AnswerVec(x) \times \text{Weight Vec} \quad (3.2.1)$$

4. 実験に用いるデータ

本実験では対象とする言語を「日本語」と「英語」をとした．日本語を用いる場合は特定分野を「医療に関する専門性」と「プログラミングに関する専門性」として，英語を用いる場合は特定分野を「プログラミングに関する専門性」としてそれぞれ実験を行う．本実験の単語の重要度計算する上で必要となるデータは，重要度の計算対象となる特定分野関連単語を抽出するために使用する辞書と単語重要度を算出するためのコーパスである．

4.1 特定分野関連単語抽出用の辞書

特定分野関連単語を抽出するために使用する辞書について説明する．

4.1.1 対象言語を日本語とした時に用いる辞書

対象言語を日本語とした時，対象とする特定分野は「医療」と「プログラミング」である．医療の関連と用語は書籍「簡潔!くすりの副作用用語事典」[12]，Wikipedia[2]，医療に関するサイト(標準病名マスター作業班[3]，看護 roo[4])から収集し，計 63,325 用語を収集した．またプログラミングの関連用語は IT 用語辞書のサイト(e-words[5])と多種多様な辞書を持つサイト Weblio[6]から情報セキュリティ用語集，OSS 用語集，NET Framework 用語集，IT 用語辞書バイナリ，コンピュータ用語辞典の計 5 種類の辞書を用いて，計 36,895 用語を収集した．

4.1.2 対象言語を英語とした時に用いる辞書

対象言語を英語とした時，対象とする特定分野は「プログラミング」である．プログラミングの関連単語は QA サイト Stack Overflow[7]に登録されている tag 名を用い，53,722 用語収集した．

4.2 単語重要度を算出するためのコーパス

単語重要度を算出するためのコーパス Dp，Dn について説明する．

4.2.1 単語重要度計算用のコーパス(日本語)

対象を日本語とした場合の実験では，Yahoo!知恵袋における「質問」と「その質問に対する回答群」をまとめて 1 つの文書として扱い，コーパスを生成した．

2 https://ja.wikipedia.org
3 http://www.dis.h.u-tokyo.ac.jp/byomei/
4 https://www.kango-roo.com/
5 http://e-words.jp/
6 http://www.weblio.jp
7 https://stackoverflow.com/

なお，本コーパスは，専門に関連する単語の重要度を求めるためのものであり，質問と回答をまとめても問題は発生しない．特定分野と一般分野の区別は質問のカテゴリを用いた．特定分野に関する質問カテゴリのついたページを特定分野のページとし，それ以外の質問カテゴリのついたページを一般分野のページとした．特定分野を医療とするときは，特定分野と判断した質問カテゴリを「病院・病気」とし，特定分野のページ数を 35,000，一般分野のページ数を 70,000 とした．一方特定分野を医療とするときは，特定分野と判断した質問カテゴリを「コンピュータテクノロジー」とし，特定分野のページ数を 15,000，一般分野のページ数を 30,000 とした．

なお，特定分野のコーパス・一般分野のコーパスは共に Mecab[13]を用いて形態素解析を行い，名詞のみを抽出した．使用した辞書は ipadic[8]に 4.1 で収集した単語を追加したものを使用した．

4.2.2 単語重要度計算用のコーパス（英語）

対象を英語とした場合の実験では，複数の Web サービスのページを用いた．特定分野の文書集合として用いたページは英語のプログラミングを専門としている QA サイト Stack Overflow[6]の質問ページ（15,000）と英語の QA サイト WikiAnswers[9]のプログラミングに関するカテゴリ「Technology」に属する質問ページ（10,000）から計 25,000 ページを収集した．一方，一般分野の文書集合としては英語のニュースサイト CNN のニュースページ[10,11]から 30,000 ページ収集し，使用した．なお，特定分野のコーパス・一般分野のコーパス内の単語は共に全て小文字に変換している．

5. 評価

本稿で提案した「ある特定分野の単語重要度を算出する手法 CrRv」の有効性を確認するための評価実験について説明する．

5.1 テストデータの用意

5.1.1 テストデータ（日本語）

対象を日本語とした時の実験では，Yahoo!知恵袋の該当特定分野に関する質問への回答の著者を対象として専門家か一般ユーザかの判定を行う．対象とする回答は，特定分野が医療の場合，カテゴリ「病院・病気」に属する質問に対しての回答である．特定分野がプログラミングの場合，カテゴリ「コンピュータテクノロジー」に属する質問に対しての回答である．

正解となる専門家（Grand Truth）は次の何れかの条件を満たすユーザとした．

1) 知恵袋内で専門家とラベルが付与されているユーザ
2) 知恵袋内でカテゴリマスターとラベルが付与されているユーザ
3) 知恵袋において回答しているユーザのうち，プロフィールから該当特定分野における専門的職業についていることが明確に判断できたユーザ

なお，3）では次の基準でユーザを選んだ．特定分野「医療」の場合，専門的職業は医者あるいは看護師であることがプロフィールから明確に判断できたユーザを対象とした．一方，特定分野「プログラミング」の場合，専門的職業はコンピュータエンジニアもしくはプログラマーであることがプロフィールから明確に判断できたユーザを対象とした．

一般ユーザは上記の条件で専門家と判断されない全ユーザとした．なお，プロフィールが空欄のユーザは本実験の対象ユーザから除外した．

5.1.2 テストデータ（英語）

対象を英語とした時の実験では，WikiAnswers の該当特定分野対象とする回答は，カテゴリ「Technology」に属する質問に対しての回答である．

正解となる専門家（Grand Truth）はプロフィールの Expert カテゴリにプログラミングに関するカテゴリが付与されているユーザとした．また，一般ユーザは専門家と判断されない全ユーザとした．なお，プロフィール上の Expert カテゴリ Interest カテゴリが双方とも空欄のユーザは本実験の対象ユーザから除外した．

5.1.3 実験に使用するデータの比率および数

表 5.1 に，各専門分野と専門家ユーザの回答数と一般ユーザの回答数を示す．表 5.1 から言語が日本語の場合，専門家ユーザの回答数と一般ユーザの回答数が約 1:4 になっていることがわかる．また，全ての実験において使用するデータセットの比率および規模は揃える必要がある．

以上から，全ての実験において使用する各データ

8 https://osdn.jp/projects/ipadic/
9 http://www.answers.com/Q/
10 http://edition.cnn.com/
11 http://money.cnn.com/

セットは専門家ユーザの回答を 100 件，一般ユーザの回答を 400 件とした.また,言語が日本語の場合はデータ数に余裕があるため，それぞれデータセットを 5 つ用意し，各データセットに対する結果の平均値をとる.

表 5.1 実際に収集した，各専門分野と専門家ユーザの回答数と一般ユーザの回答数

対象分野と言語	専門家ユーザの回答数	一般ユーザの回答数
医療(日本語)	12,302	41,058
プログラミング(日本語)	1,302	4,093
プログラミング(英語)	122	1,618

5.1.4 ベースライン手法

提案手法 CrRv の比較対象(ベースライン)として，既存の 4 手法（2.1 で示した tf-idf と，2.2 で示した tf-rf，tf-idfec_b, tf-PNF[2])を用いる.さらに，提案手法では既存手法とは異なり tf 値を用いていないことから，tf 値を用いる妥当性も同時に評価するため，既存手法 CrRv に tf 値を掛け合わせた tf-CrRv との比較も同時に行う.

tf-idf を用いた専門辞書作成では，提案手法で使用した特定分野のコーパス Dp のみを使用した.今回の重みづけは当該特定分野にどれだけ精通しているかを判断できることを目的としているため，一般分野のコーパス Dn は用いない．単語 t のドキュメント d∈Dp に対する重要度 w(t,d)の計算には，式(5.1)を用いる.

$$W(t) = \max_{d \in Dp} w(t,d) \qquad (5.1)$$

tf-rf, tf-idfec_b, tf-PNF[2] を用いた専門辞書の作成では，提案手法と同様に種類のコーパス Dp, Dn を使用する.

5.1.5 QA サイトの回答者の専門性推定手法

本実験では，ある回答に対し，その著者が専門家か否かで推定を行い評価する．まず，推定対象となる全ての回答に対して専門性スコアを計算し，付与する．その後専門性スコアでランキングを生成し，Precision@k で評価する．本研究の目的は，短い文章からいかに専門性を判断できるかどうかにあるため，文章長を可変させながら手法の評価を行う．これを実現するため，推定対象となる回答単位で専門性スコア(x)を計算するのではなく,全ユーザの各回答の先頭の n 文字までを切り出した x[:n]を用いて 3.2 で述べた AnswerScore(x[:n])を計算する.

続いて，tf-CrRv およびベースライン手法への適用方法について説明する．ベースライン手法の多くは提案手法 CrRv(t)とは違い，tf(t,x) を用いている．そのため，3.2 で述べた AnswerVec をそのまま用いるのではなく，AnswerVecTF(x)=[tf(x, t_1),tf(x,t_2) … tf(x,t_j)，…tf(x, $t_{|T|}$)]を用いる．式(3.2.1)で示した AnswerScore 内の式 AnswerVec を AnswerVecTF に変えた式 AnswerScoreTF(x)を式(5.3.1)に示す.

$$AnswerScoreTF(x) = AnswerVecTF(x) \times \text{Weight Vec} \qquad (5.3.1)$$

6. 実験結果

対象言語が日本語の場合，専門家の回答を 500 件，一般人の回答を 2,000 件用い，これを 5 つのデータセットに排他的に分割し実験を行ない，各データセットに対して評価を行いその平均値をとった.

対象言語が英語の場合，専門家の回答を 100 件，一般人の回答を 400 件とした．データセットの数の選定の基準は日本語の場合と同様にするためである.なお，英語のデータセットでは，収集したデータ集合がデータセットを排他的に分けるに十分なデータが収集できなかったためデータセットは 1 つとした.で行なった.

なお,言語を問わず,対象は長さが 140 文字以上(英語の場合も文字数）の回答とし,使用文字数は 10 から 140 まで 10 文字ずつ変化させ，実験を行なった．評価は Precision@10 を用いた.

結果を図 6.1～図 6.3 及び表 6.1～表 6.3 に示す．表 6.1～表 6.3 では，手法ごとの推定結果の最大値とその最大値を得た時の使用文字数をまとめている。

6.1 特定分野・医療及びプログラミング(言語:日本語)の結果

表 6.1,表 6.2 と図 6.1,図 6.2 から全体的に提案手法である CrRv および tf-CrRv の精度が高い結果となった．既存の単語重要度計算手法の目的が専門性推定ではなく文書のカテゴリ分類であることから，結果は妥当と言える.

分野ごとにみるとコンピュータ分野に比べて医療分野の精度が低い．理由として，質問者の専門性レベルの違いが考えられる.実験では Yahoo!知恵袋を用いており，分野をプログラミングとした時は「コンピュータテクノロジー」カテゴリに投稿された質問に対する回答を対象としている.「コンピュータテクノロジー」カテゴリには専門的な質問が比較的多く存在するため，回答も専門的な回答が多い．そのため専門用語の出現回数が多かったと考えられる.

一方,「病院・病気」カテゴリには一般の人の質問の投稿も多く存在する．そのため専門家も一般の人のも

わかるような単語のみを用いて回答を行うことが多い．したがって，プログラミング分野に比べて一般人が知りえない専門用語の出現回数が少なかったことが原因と考えらえる．

6.2 特定分野・プログラミング(言語:英語)の結果

表 6.3，図 6.3 から全体的に CrRv の精度が高い結果となった．しかし，Yahoo!知恵袋を使用した場合に比べて低い．理由として以下の三点が考えられる．

- 今回，「install」と「installing」のような英単語の変化に対して対応できていない．このため，これらの単語が別単語として判断されたこと．
- 特定分野のコーパス Dp に WikiAnswers だけでなく別サイトである StackOverflow を用いており，適応先である WikiAnswers とコーパスとしての専門性が異なったこと．
- WikiAnswers をデータセットとした際に，専門用語を StackOverflow のタグ名と定義したが，同タグの中には，プログラムでは多用されるが日常でも多用される用語(「main」「for」「if」など)も含まれる．

以上の理由から，日本語での実験結果（Yahoo!知恵袋を使用した場合）に比較し適切に重みを付与できなかった可能性がある．

7. おわりに

本稿では，短い文章における著者の特定分野の精通度合いを判断することを目的とした「単語重要度計算手法 CrRv」を提案した．特定分野への精通度合いを判断することを目的としているため，提案手法では該当分野に精通していないと知り得ない単語に高い重要度を付与する．評価実験の結果，既存手法である tf-rf，tf-PNF2，tf-idfec_b と比較して Precision@10 で日本語の場合 0.2，英語の場合で 0.3 の向上を確認した．

今後の課題としてはさらなる精度向上，重要度を付与した後の対象文書の専門性レベルの計算方法の再考，他の分野への適用などが考えられる．

表 6.1 特定分野・医療(言語:日本語)の時の Precision@10 の最大値とその時使用した文字数

手法名	CrRv(提案手法)	tf-CrRv(提案手法)	tf-rf	tf-idfec_b	tf-PNF2	tf-idf
最大値	**0.56**	**0.56**	0.46	0.46	0.48	0.40
使用文字数	10	10	40	10	20	80

表 6.2 特定分野・プログラミング(言語:日本語)の時の Precision@10 の最大値とその時使用した文字数

手法名	CrRv(提案手法)	tf-CrRv(提案手法)	tf-rf	tf-idfec_b	tf-PNF2	tf-idf
最大値	**0.70**	0.63	0.52	0.39	0.36	0.25
使用文字数	30	30	20	20	10	20

表 6.3 特定分野・プログラミング(言語:英語)の時の Precision@10 の最大値とその時使用した文字数

手法名	CrRv(提案手法)	tf-CrRv(提案手法)	tf-rf	tf-idfec_b	tf-PNF2	tf-idf
最大値	**0.50**	**0.50**	0.30	0.20	0.30	0.30
使用文字数	70	70	60	10	60	20

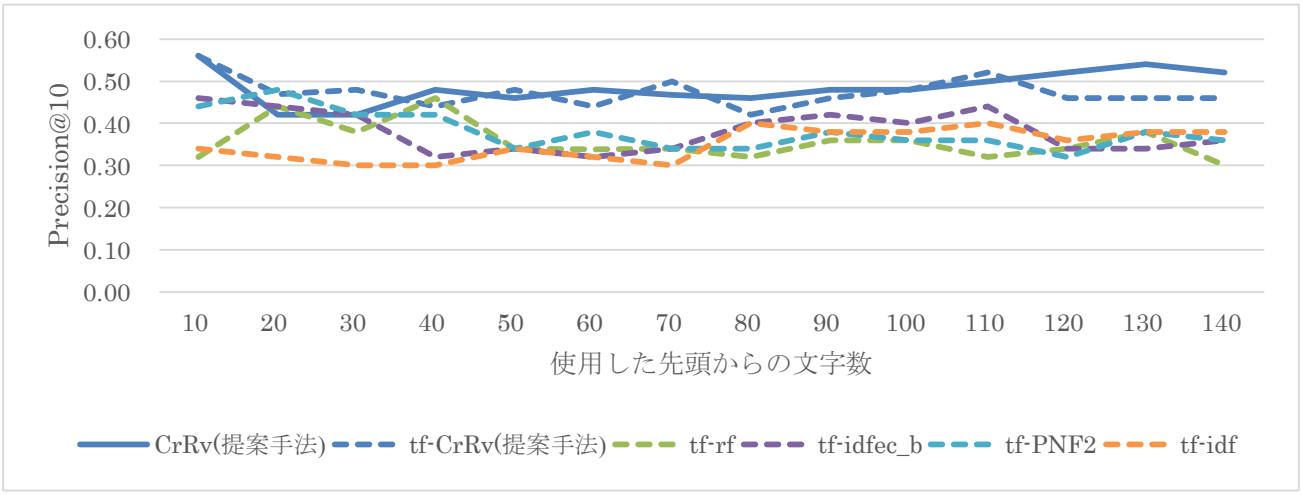

図 1 特定分野・医療(言語:日本語)の時の Precision@10

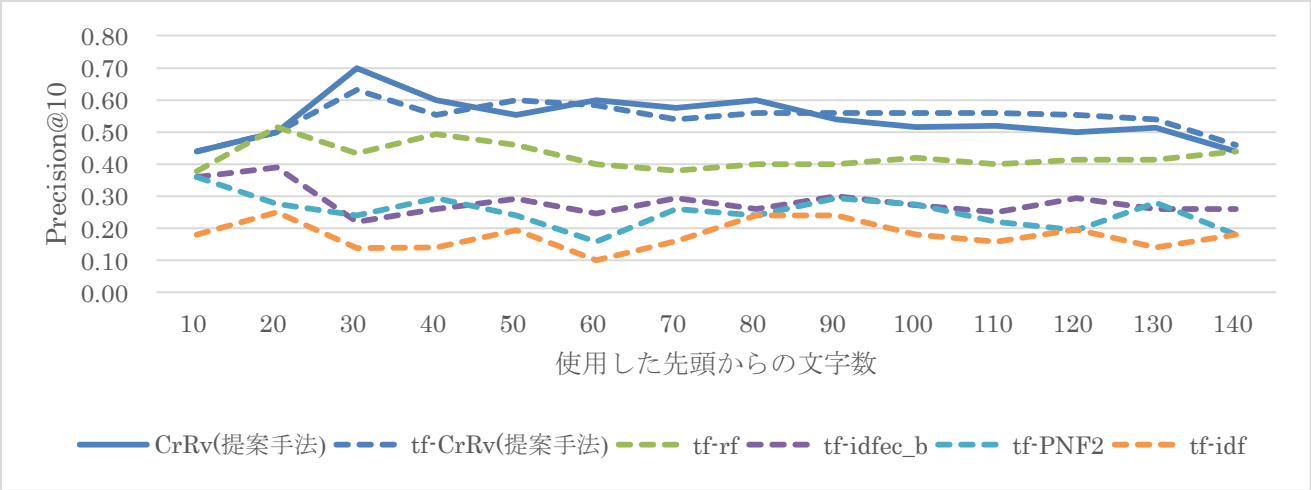

図 2 特定分野・プログラミング(言語:日本語)の時の Precision@10

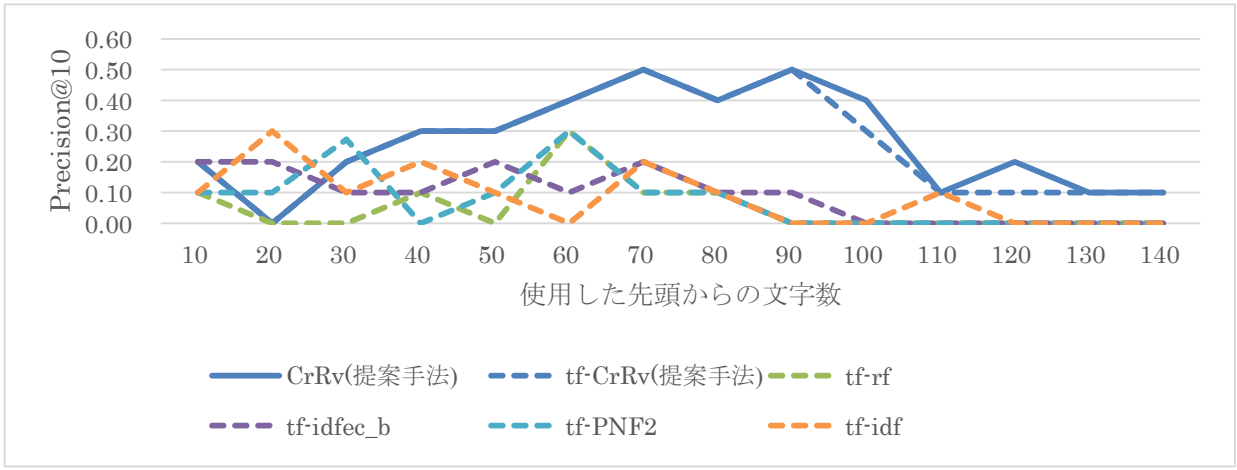

図 3 特定分野を・プログラミング(言語:英語)の時の Precision@10

参 考 文 献

[1] Iyyer, M., Boyd-Graber, J. L., Claudino, L. M.B., Socher, R.,& Daumé III, H. "A Neural Network for Factoid Question Answering over Paragraphs, " *EMNLP* , pp.633-644 (2014)

[2] Munger, Tyler, and Jiabin Zhao. "Identifying influential users in on-line support forums using topical expertise and social network analysis." Proc. of *2015 IEEE/ACM Int'l Conf. on Advances in Social Networks Analysis and Mining (ASONAM)* (2015).

[3] Lim, Wern Han, Mark James Carman, and Sze-Meng Jojo Wong. "Estimating Domain-Specific User Expertise for Answer Retrieval in Community Question-Answering Platforms." *Proc. of the 21st Australasian Document Computing Symposium*. ACM, pp.33-40, (2016).

[4] 池田和史, 服部元,松本一則. "マーケット分析のための twitter 投稿者プロフィール推定手法", 情処論 (CDS),Vol .2, No.1, pp.82-93 (2012)

[5] X.Shao, Z.Chunhong and J.Yang. "Finding Domain Experts in MiCroblogs" Proc. of the 10th Int'l Conf. on WEBIST (2014).

[6] Yang, Zichao, et al. "Hierarchical Attention Networks for Document Classification." Proc. of NAACL HLT 2016 (2016).

[7] 滝川真弘, 山名早人.“特定分野における単語重要度計算手法の提案と短い文章における著者の専門性推定への適応”, 情処研報(NL),Vol. 2017-NL-233, No.6, pp.1-6 (2017)

[8] G.Saltion, E.A.Fox and H.Wu. "Extended Boolean Information Retrieval," CACM, Vol.26, No.11, pp.1022-1036 (1983).

[9] M. Lan, C.L.Tan, J.Su and Y.Lu. "Supervised and traditional term weighting methods for automatic text categorization" IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.31, No.4, pp.721-735 (2009).

[10] Naderalvojoud, Behzad, Ebru Akcapinar Sezer, and Alaettin Ucan. "Imbalanced text categorization based on positive and negative term weighting approach." TSD 2015. LNCS, Vol. 9302. Springer (2015)

[11] Domeniconi, Giacomo, et al. "A Study on Term Weighting for Text Categorization: A Novel Supervised Variant of tf.idf." Proc. of the 4th Int'l Conf. on Data Management Technologies and Applications(DATA2015), pp.26-37(2015)

[12] くすりの適正使用協議会, 簡潔!くすりの副作用用語事典, pp.1-356, 第一メディカル (2003)

[13] T.Kudo, K.Yamamoto and Y.Matsumoto. "Applying Conditional Random Fields to Japanese Morphological Analysis," Proc. of the 2004 Conf. on EMNLP, pp.230-237 (2004)