Vol.2017-NL-233 No.15
2017/10/25

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

滝川真弘[†1]　山名早人[†1]

概要：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．短い文章とは，例えば質問投稿サイトの回答などが挙げられる．こうした短い文章単体から著者の専門性を判定する場合，特徴量が不足するため，既存研究では当該著者により記述された複数の文章（あるいは他の属性）を用いた推定を行っている．しかし，当該著者に対して常に複数の文書が用意できるとは限らない．この問題を解決するため，本研究では，出現する単語に専門毎に適切な重みを付与し，著者の専門性を短い文章からも推定できる手法を提案する．具体的には、単語の重み付与手法として CrRv を提案する．評価実験においては，データセットを Yahoo!知恵袋，対象特定分野を「医療」と「プログラミング」として回答者の専門性の推定を行った．Precision@10 で評価したところ,医療分野においては 0.56, プログラミング分野においては 0.70 となり，既存手法である tf*rf, tf*PNF^2, tf*idfec-b と比べて有用性を確認した．

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る．いずれも，ある特定分野における著者の専門性は，投稿の信頼性などに対して重要な要素である．しかし,短い文章は一般的な文書とは異なり文量が小さいため，情報量も小さくなってしまう．具体的には出現する単語の種類や単語数が小さくなる．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].

そこで，既存手法の中では機械学習を使わず，他の情報を用いて少ない情報量を補い，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用いるには，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，機械学習を用いるには十分な学習データが必要となる．Yang らは学習のために 24 万から 240 万のデータを使用していることから分かる通り，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，文章内の単語自体に専門分野別の重み（重要度）を付与する手法を提案しており，内部で 2 つのハイパーパラメータを用いている．しかし，ハイパーパラメータ決定のためには事前実験が必要となり，大きなデータセットの準備が必要となる．そこで，本稿では新たにハイパーパラメータを用いない手法を提案する．さらに，評価対象のデータを２つに増やし詳細な実験を行う．

本稿では,「適切な重みは対象とする分野毎に異なる」ことを前提に「特定分野を対象とした単語重要度の計算法」について提案する．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．

†1 早稲田大学
Waseda University

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． x

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(2.1.2)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(2.1.3)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 の出現が，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(2.9)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(2.2.8)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

本稿では，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する．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

†1 早稲田大学
Waseda University

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．

4. 実験に用いるデータ

本節では，実験に用いるデータについて述べる．今回の実験では対象とする特定分野を「医療に関する専門性」と「プログラミングに関する専門性」として実験を行う．

4.1 特定分野関連単語を抽出するために使用する辞書

医療の関連単語として，書籍「簡潔!くすりの副作用用語

事典」[12]と Wikipedia(a)，それから医療に関するサイトである標準病名マスター作業班(b)，看護 roo(c)から関連用語を収集し 63,325 語収集した．また，プログラミングの関連単語として，IT 用語辞書のサイトである e-words(d)と多種多様な辞書を持つサイトである Weblio(e)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 種類の辞書を利用し，のべ 36,895 の専門用語（単語）を収集した．本辞書に出現する単語を対象に 4.2 項のコーパスを用いて単語重要度を付与する．

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．本実験では，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．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．

対象を医療分野とした際は質問のカテゴリが「病院・病気」となっているものを特定分野の文章とし，35,000 ページを使用した．また．それ以外を一般文書の文章として扱い，70,000 ページを使用した． 一方，対象をプログラミングとした際は，質問のカテゴリが「コンピュータテクノロジー」となっているものを特定分野の文章とし 15,000 ページを使用した．また，それ以外を一般文書の文章として扱い，30,000 ページを使用した．

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(f)に 4.1 節で収集した単語を追加したものを使用した．

5. 評価方法

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」の有効性を確認するため，Yahoo!知恵袋の該当特定分野に関する質問への回答の著者が専門家か一般ユーザかで評価を行う．正解となる専門家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) 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なユーザ

また，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 ベースライン手法

提案手法の比較対象（ベースライン）として，既存の 4 手法（2.1 項で示した TF-IDF と，2.2 項で示した tf*rf, tf*idfec-b，tf*PNF^2）を用いる．

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．

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

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．

5.2 Yahoo!知恵袋の回答者の専門性の定量的推定手法

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．そこで，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．その後スコアでランキングを生成し，Precision@k で評価する．しかし，回答そのもので専門性スコアを計算すると，回答ごとの文書長の影響を受けてしまうため公平にスコアの計算ができない．そこで，スコアの計算を行う際は先頭 n 文字までを用いて計算を行う．この考え方は，ベースライン手法にも適用する．

5.2.1 回答の専門性スコア計算方法

ある回答 x の先頭 n 文字目までを用いた際の専門性スコアを AnswerScore(x, n)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≤j≤|T|)が回答 x の中で出現した回数を TF ($t_{j,\,x,\,n}$)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，n)=[TF(x,n, t_1)，TF(x,n, t_2)，…TF(x,n, t_j)，…TF(x,n, $t_{|T|}$)]，|T|次元の単語重要度ベクトルを WeightVec=[W(t_1)，W(t_2)，…W(t_j)，…W($t_{|T|}$)]とした時，AnswerScore(x, n)を式(5. 2)に示す．この際，CrRv は既存研究とは異なり対象文書の tf 値は考慮しないため，$TF_{x,\,n}(t_j)$は，値が 1 以上の場合は全て 1 とした．

$$AnswerScore(x,n) = AnswerVec(x,n) \times \text{Weight Vec} \tag{5.2}$$

6. 実験結果

6.1 特定分野を医療とした時の結果

専門家の回答を 500 件，一般人の回答を 2000 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，

a https://ja.wikipedia.org
b http://www.dis.h.u-tokyo.ac.jp/byomei/
c https://www.kango-roo.com/
d http://e-words.jp/
e http://www.weblio.jp
f https://osdn.jp/projects/ipadic/

表 1 医療分野を特定分野とした時の各手法における Precision@10 の値の最大値とその時の使用文字数

	CrRv	tf*rf	tf*idfec-b	tf*PNF^2	tf*idf
最大値 (Precision@10)	0.56	0.48	0.44	0.46	0.40
使用文字数	10	20	90	20	100

表 2 プログラミング分野を特定分野とした時の各手法における Precision@10 の値の最大値とその時の使用文字数

	CrRv	tf*rf	tf*idfec-b	tf*PNF^2	tf*idf
最大値 (Precision@10)	0.70	0.50	0.39	0.36	0.28
使用文字数	30	20	20	10	90

それぞれで評価を行いその平均値をとった．それぞれのデータセットは，専門家の回答を 100 件，一般人の回答を 400 件である．対象とする回答は，カテゴリが「病院・病気」に属する質問に対しての回答である．使用文字数は 10 から 200 まで 10 文字ずつあげ，それぞれ実験を行なった．評価は Precision@10 を用いた．結果を図 1 に示す．また，手法ごとの推定結果の最大値と最大値を出した時の使用文字数を表 1 にまとめる．なお，対象は長さが 200 文字以上の回答とした．図 1, 表 1 から使用文字数が 10 の時の CrRv の結果が最も高いことがわかる．

6.2 特定分野をプログラミングとした時の結果

専門家の回答を 500 件，一般人の回答を 2000 件とし，排他的に 5 つのデータセット（専門家の回答 100 件，一般人の回答 400 件）に分けて実験を行ない，それぞれで評価を行いその平均値をとった．対象とする回答は，カテゴリが「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．使用文字数は 10 から 200 まで 10 文字ずつあげ，それぞれ実験を行なった．評価は Precision@10 を用いた．結果を図 1 に示す．また，手法ごとの推定結果の最大値と最大値を出した時の使用文字数を表 1 にまとめる．となお，対象は長さが 100 文字以上の回答とした．図 2，表 2 から使用文字数が 30 の時 CrRv の結果が最も高いことがわかる．また，図 2 から常に CrRv が他の手法に比べ精度が高いことがわかる．

6.3 考察

図 1, 図 2 から全体的に 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．一方，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

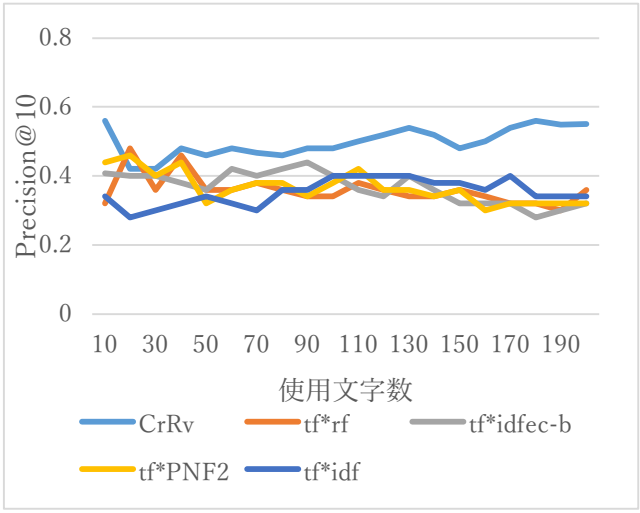

図 1 医療分野を特定分野とした時の各手法におけるそれぞれの文字数を用いた際の Precision@10 の値

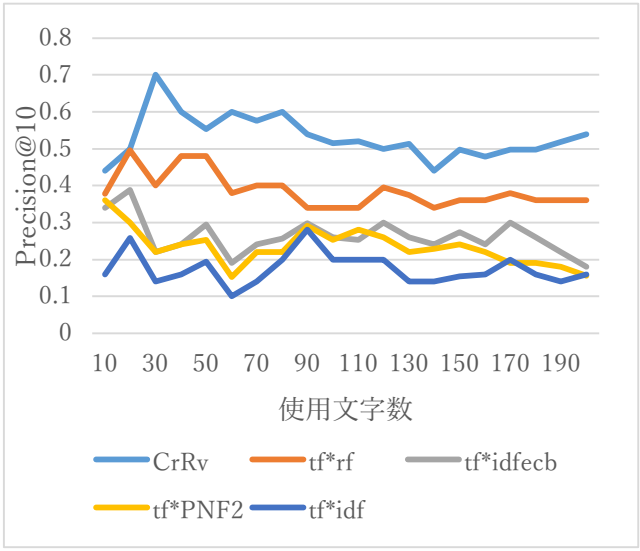

図 2 コンピュータ分野を特定分野とした時の各手法におけるそれぞれの文字数を用いた際の Precision@10 の値

野をコンピュータとした時は「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，コンピュータ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.

また，表 1，表 2，図 1，図 2 から対象回答の専門性レベルを計算するために使用する文字数の長さと精度に相関がないことがわかる．特に，コンピュータ分野においては使用文字数が少ない時の方が，全体的に精度が高い．今回の実験では対象回答の専門性レベルを計算する際，出現するすべての専門用語の重要度を合計して計算している．そのため重要度の低い専門用語が多く出現する回答の専門性レベルは高く計算されてしまう．算出した重要度を用いた回答の専門性レベルの計算方法は今後の課題である.

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．評価実験においては，データセットを Yahoo!知恵袋，対象特定分野を医療とコンピュータとして回答者の専門性の推定を行った。Precision@10 で評価したところ, 医療分野においては CrRv が 0.56, コンピュータ分野においては CrRv が 0.70 の精度となり, 医療分野では既存手法と比べて 0.08，プログラミング分野では 0.20 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.

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." *Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference on*. IEEE, (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." *Proceedings 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", 情報処理学会論文誌コンシューマ・デバイス&システム (CDS),Vol .2, No.1, pp.82-93 (2012)

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

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

[7] 滝川真弘，山名早人. “特定分野を対象とした単語 重要度計算手法の提案と Twitter における専門 性推定への適応”, FIT2016(第 15 回情報科学技術 フォーラム)，第 2 分冊，pp.1-7 (2016)

[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. Lecture Notes in Computer Science, vol 9302. Springer, Cham(2015)

[11] Domeniconi, Giacomo, et al. "A Study on Term Weighting for Text Categorization: A Novel Supervised Variant of tf. idf." DATA 2015 Proceedings of 4th International Conference on Data Management Technologies and Applications 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, pp1-356, 第一メディカル,2003/9

[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).