2017 年度 修士論文

特定分野における単語重要度計算手法の提案と

短文からの著者専門性推定への適応

提出日: 2018 年 1 月 30 日
指導: 山名 早人 教授

早稲田大学大学院 基幹理工学研究科
情報理工・情報通信専攻
学籍番号:5116F056-4

滝川 真弘

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.こうした短い文章が内包する特徴量は少ないため，既存研究では当該著者により記述された複数の文章や他の属性を用いることで特徴量を増やし，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ことを目的とする方法「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. precision@10 で評価を行い，既存手法である tf-rf, tf-PNF2, tf-idfec_b と比較実験を行なった. 結果として，既存手法の内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.

目次

第1章　　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，推定する手法をとっている．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．

本論文で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著者の専門性を推定する手法に取り組む．具体的には，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 CrRv を提案する．

本稿では，「単語に付与する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 CrRv では，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ることを前提とし，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，当該分野コーパス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，第 2 章にて関連研究，第 3 章にて提案手法．第 4 章にて実験に使用するデータセット，第 5 章にて評価方法，第 6 章にて実験結果を示し，第 7 章にて本稿をまとめる．

第2章　　関連研究

第2章では関連研究について述べる. 具体的には, 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. 筆者は,関連研究は大きく二つに分類できると考えた.2.1 にて文章において単語重要性を図る手法, 2.2 においてカテゴリと単語の関係から重要度を計算する手法について述べる.　2.3 にて, 関連研究のメリットとデメリットをまとめる.

2.1　文章において単語重要性を測る手法

2.1 では文章において単語重要性を図る手法について説明する. 2.1.1 において手法 tf-idf[8]について説明する.

2.1.1 tf-idf[8]

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(2.1.2)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. $idf(\mathrm{t})$は, 式(2.1.2)により計算する.

$$w(t,d) = tf(t) * idf(t) \tag{2.1.1}$$

$$tf(t) = \log\left(1 + count(t,d)\right) \tag{2.1.2}$$

$$idf(t) = log\left(\frac{|D|}{df(t)}\right) \tag{2.1.3}$$

ここで, count(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 w(t,d)を算出することにより, 対象とする文書 d の特徴量を求めている. このため, ある分野に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. この理由は tf-idf 自体が文書 d に対する単語 t の重みを求めるものであり, ある文書集合に対する単語 t の重みを求める手法ではないためである. tf-idf を用いて特定分野の文書集合 D に対する単語重要度 W(t,D)を求めるには$W(t,D) = \sum_{d \in D} w(t,d)$を適用することができる. しかし,

特定分野の文書集合 D には,当該特定分野に関連しない単語も一般的に含まれるため，当該特定分野に関連しない単語に大きな重みが付与される可能性がある.

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，2.2.1 において tf-rf[9]，2. 2. 2 において tf-PNF^2[10]，2.2.3 において tf-idfec_b[11]をそれぞれ紹介する．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， Dp 内で t が出現する文書数を$n_{Dp}(t)$， Dp のうち t が出現しない文書数を$|D_p| - n_{Dp}(t)$， Dn のうち t が出現する文書数を$n_{Dn}(t)$， Dn のうち t が出現しない文書数を$|D_n| - n_{Dn}(t)$，全文書数を$|D_p| + |D_n|$とする．

2.2.1 tf-rf[9]

2009 年 East China Normal University にの 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 d 内での単語出現頻度 tf(t,d)に加え，単語 t の出現が，カテゴリ C に属する文書集合 Dp と当該カテゴリに属さない文書集合 Dn でどれだけ異なるか，すなわち$n_{Dp}(t)$と$n_{Dn}(t)$の比率を用いている．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．

$$rf(t) = \log\left(2 + \frac{n_{Dp}(t)}{\max(1, n_{Dn}(t))}\right) \tag{2.2.2}$$

なお，tf は文書 d 中の単語 t の出現頻度であり，2.1.1 で説明した 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．

2.2.2 tf-PNF^2[10]

2015 年に Hacettepe University の 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，カテゴリ C に属する文書集合 Dp と属さない文書集合 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 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 は文書 d 中の単語 t の出現頻度であり，2.1.1 で説明した tf-idf の tf と同値である．tf-PNF^2 は，tf(t, d)と PNF^2 (t)の積により求める．

$$PNF^2(t) = \frac{P(t_i \mid Dp) - P(t_i \mid Dn)}{P(t_i \mid Dp) + P(t_i \mid Dn)} \tag{2.2.3}$$

$$P(t_i \mid Dp) = \frac{n_{Dp}(t)}{|D_p|} \tag{2.2.4}$$

$$P(t_i \mid Dn) = \frac{n_{Dn}(t)}{|D_n|} \tag{2.2.5}$$

2.2.3 tf-idfec_b[11]

University of Bologna の Giacomo ら[11]は 2015 年に tf-idfec_b(t)[1]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(2.2.6)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.

$$idfec\_b(t) = \log\left(2 + \frac{n_{Dp}(t) + |Dn|}{\max(1, n_{Dn}(t))}\right) \tag{2.2.6}$$

2.3 まとめ

2.1 で述べた手法は検索インデックスの作成のためであり, 2.2 で述べた手法はいずれも未知の文章のカテゴライズを目的としている. したがって, どちらも本論文の目的とする専門性の高い単語への重要度を付与に用いることは出来ないと考えられる. 第2章で説明した各手法を表 2.1 にてまとめる.

表 2.1 関連研究のまとめ

手法名	計算式	目的	重要視する点	提案年
tf-idf[8]	$\log(1 + tf(t,d)) * log\left(\frac{\|D\|}{df(t)}\right)$	インデックス付与	特定の文書のみに出現する傾向	1983 年
tf-rf[9]	$\log\left(1 + tf(t,d)\right) * \log\left(2 + \frac{n_{Dp}(t)}{\max(1, n_{Dn}(t))}\right)$	文書のカテゴライズ	$n_{Dp}(t)$と$n_{Dn}(t)$の比率	2009 年
tf-PNF[2][10]	$\log\left(1 + tf(t,d)\right) * \frac{P(t_i \mid Dp) - P(t_i \mid Dn)}{P(t_i \mid Dp) + P(t_i \mid Dn)}$	文書のカテゴライズ	Dp, Dn 内で単語 t が出現する確率	2015 年
tf-idfec_b[11]	$\log\left(1 + tf(t,d)\right) * \log\left(2 + \frac{n_{Dp}(t) + \|Dn\|}{\max(1, n_{Dn}(t))}\right)$	文書のカテゴライズ	単語 t が如何に該当カテゴリ以外で出現しないか	2015 年

1 元々の論文では idfec_b だが, マイナスに見えてしまうため本論文では idfec_b と表記する

第3章　　CrRv(t)手法の提案

本章では，提案手法について述べる．CrRv(t)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」をである．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，3.1にて詳細を述べる．

3.1　CrRv(t)

提案する CrRv(t)の概要図を図 3.1 にて示し，CrRv(t)を求める計算式を式(3.1.1)に示す．その後，式(3.1.1)に示した各項についての詳細を 3.1.1，3.1.2，3.1.3，3.1.4 にて説明する．

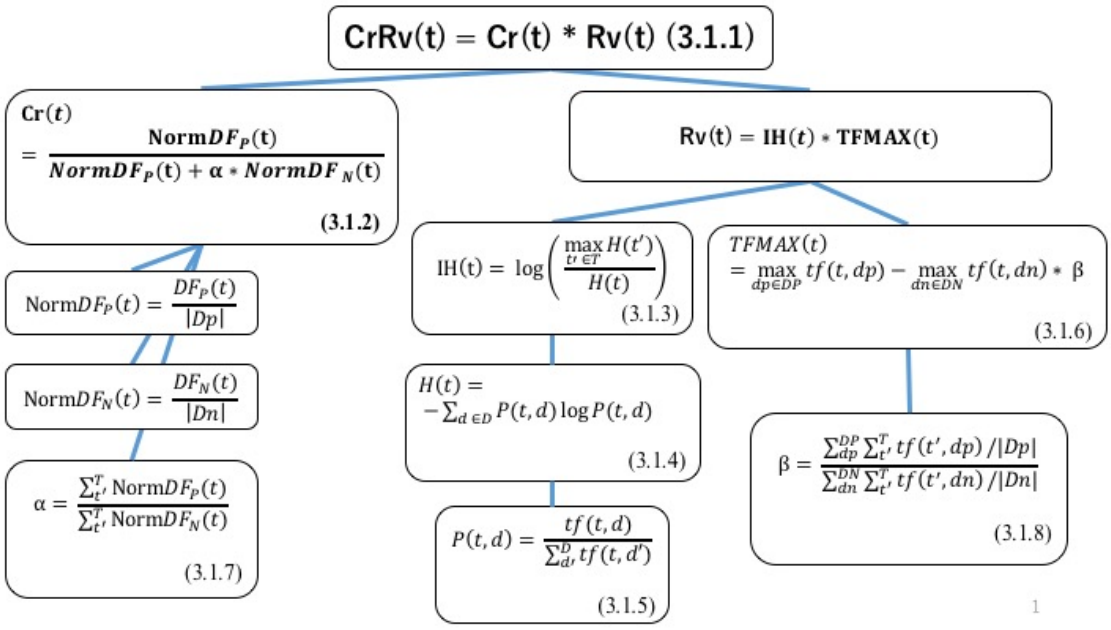

図 3.1 提案手法の概要図

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

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

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

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

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

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

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

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

また，上式において使用する変数を表 3.1 にて示す．

表 3.1 上式において使用する変数とその説明

対象とする単語	t
特定分野コーパス	Dp
一般分野コーパス	Dn
全文書集合	D（=Dp+Dn）
全単語集合	T
Dp の文書の数	|Dp|
文書 d 中に出現する単語 t の数	tf(t,d)
単語 t の Dp における文書出現頻度	$n_{Dp}(t)$
単語 t の Dn における文書出現頻度	$n_{Dn}(t)$
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	α，β

3.1.1 Cr(t): 特定分野コーパス Dp への出現頻度の偏り

Cr(t)は単語 t の Dp の出現頻度の偏り具合を示し，Dp への片寄りが強い単語に大きな重要度を付与する．これは，前述した「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」に基づくものである．ただし，|Dp|と|Dn|は同一ではないため正規化している．一方，α は$n_{Dn}(t)$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．

3.1.2 IH(t): 単語 t の特異性

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．具体的に述べると，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）である．すなわち，「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」なるため，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．

3.1.3 TFMAX(t): ノイズとなる単語の重みの減少

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．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$\max_{d\in D} tf(t,d)$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$\max_{d\in D} tf(t,d)$から$\max_{d\in D} tf(t,d)$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，$\max_{d\in D} tf(t,d)$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．

3.1.4 パラメータ決定法

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．また，より詳細な説明をそれぞれ 3.1.4.1，3.1.4.2 にて説明する．

3.1.4.1 α：Cr(t)で使用するパラメータ

パラメータ α は Cr(t)において$n_{Dn}(t)$の影響を調整するパラメータである．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．

3.1.4.2　　β：TFMAX(t)で使用するパラメータ

β は項 TFMAX(t)において，$\max_{d \in D} tf(t,d)$の影響を調整するためのものである．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 を用いた文章からの著者専門性計算の流れ

本節では提案手法 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．概要図を図 3.2 に示す．

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≤j≤|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(t_j), …W($t_{|T|}$)]とした時，AnswerScore(x)を式(3.2.1)に示す．

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

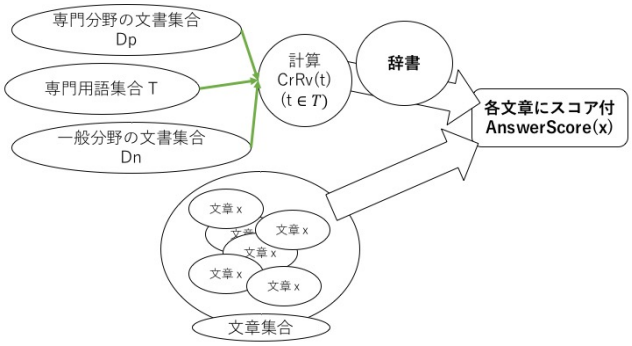

図 3.2 文章からの著者専門性計算の流れの概要図

第4章　　実験に用いるデータ

本章では，実験に用いるデータについて述べる．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．特定分野を「医療」および「プログラミング」とした理由は，どちらの分野も関連用語が書籍や Web サイトを用いることで容易に収集できることと，関連用語の中に非専門家が使用する単語(風邪やファイル)が多く存在するためである．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 にて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述べる．4.2 にて単語重要度を算出するためのコーパスについて述べる．

4.1　　特定分野関連単語を抽出するために使用する辞書

4.1 では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 では対象言語を日本語とした時の実験で使用する辞書について，4.1.2 では対象言語を英語とした時の実験で使用する辞書についてそれぞれ説明する．

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．それぞれの分野における関連用語の収集先と収集した数を表 4.1 にてまとめる．本辞書に出現する単語を対象に 4.2.1 のコーパスを用いて単語重要度を付与する．

表 4.1 対象言語を日本語とした時の，特定分野とそれぞれの分野の関連用語の収集先および収集数

分野	収集先	収集数
医療	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])	63,325
プログラミング	IT 用語辞書のサイト(e-words[5])	36,895

[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/

	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	

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の収集先および収集数を表 4.2 にてまとめる．本辞書に出現する単語を対象に 4.2.2 のコーパスを用いて単語重要度を付与する．

表 4.2 対象言語を英語とした時の，特定分野と
それぞれの分野の関連用語の収集先および収集数

分野	収集先	収集数
プログラミング	QA サイト Stack Overflow[7]に登録されている tag 名	53,722

4.2 単語重要度を算出するためのコーパス

4.2 では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対象言語が日本語の時の単語重要度計算のためのコーパス

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．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．表 4.3 に収集したカテゴリと収集数をまとめる．

表 4.3 対象言語を日本語とした時の，特定分野と
Dp,Dn の収集先および収集数

特定分野	特定分野と判断した質問カテゴリ	特定分野のページ数 (Dp)	一般分野のページ数 (Dn)
医療	病院・病気	35,000	70,000
プログラミング	コンピュータテクノロジー	15,000	30,000

6 http://www.weblio.jp

7 https://stackoverflow.com/

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対象言語が英語の時の単語重要度計算のためのコーパス

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． 表 4.4 に Dp,Dn の収集先および収集についてまとめる．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．

表 4.4 対象言語が英語，対象特定分野をプログラミングとした時の，Dp,Dn の収集先と収集数

	収集先	
特定分野 (Dp)	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 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	25,000 (StackOverflow から 15,000, WikiAnswers から 10,000)
一般分野 (Dn)	英語のニュースサイト CNN のニュースページ[10,11]	30,000

8 https://osdn.jp/projects/ipadic/

9 http://www.answers.com/Q/

10 http://edition.cnn.com/

11 http://money.cnn.com/

第5章　　評価

第5章では，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実験について説明する．5.1 において使用する実験データとラベル付けの方法について説明する．5.2 にて比較対象となるベースライン手法について述べる.5.3 にて回答者の専門性の定量的推定手法について述べる．

5.1　テストデータの用意

5.1 では使用する実験データとラベル付けの方法について説明する．5.1.1 において対象言語を日本語とした時のテストデータについて説明する．5.1.2 において対象言語を英語とした時の実験に使用するデータについて説明する．5.1.3 において，実験に使用するデータの比率および数について説明する．

5.1.1 対象を日本語とした時に使用するテストデータ

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．

5.1.1.1　専門家ユーザ

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．特定分野を「医療」とした場合，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のを対象とした．一方，特定分野を「プログラミング」とした場合，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のを対象とした．

なお，1) 専門家とラベルが付与されているユーザと 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の実際の画面キャプチャをそれぞれ図 5.1，図 5.2 にて示す．

図 5.1 Yahoo!知恵袋上で実際に専門家ラベルが付与されているユーザの
プロフィール画面キャプチャ[12]

図 5.2 Yahoo!知恵袋上で実際にカテゴリマスターラベルが付与されているユーザの
プロフィール画面キャプチャ[13]

[12] https://chiebukuro.yahoo.co.jp/my/yc_allabout_tomkurata

[13] https://chiebukuro.yahoo.co.jp/my/gsytd341

5.1.1.2　一般ユーザ

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，5.1.1.1 で示した条件の 1)，2)が満たされず，かつ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 対象を英語とした時に使用するテストデータ

対象を英語とした時の実験では，WikiAnswers 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．Wikianswers には自分のプロフィールに Expert カテゴリと Interest カテゴリを選択することができる．

5.1.2.1　専門家ユーザ

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．

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

プログラミングに関するカテゴリとは，WikiAnswers 上の「Technology」カテゴリおよび「Technology」カテゴリのサブカテゴリ全般としている．実際の専門家ユーザと定義したユーザの例の画面キャプチャを図 5.3 に示す．

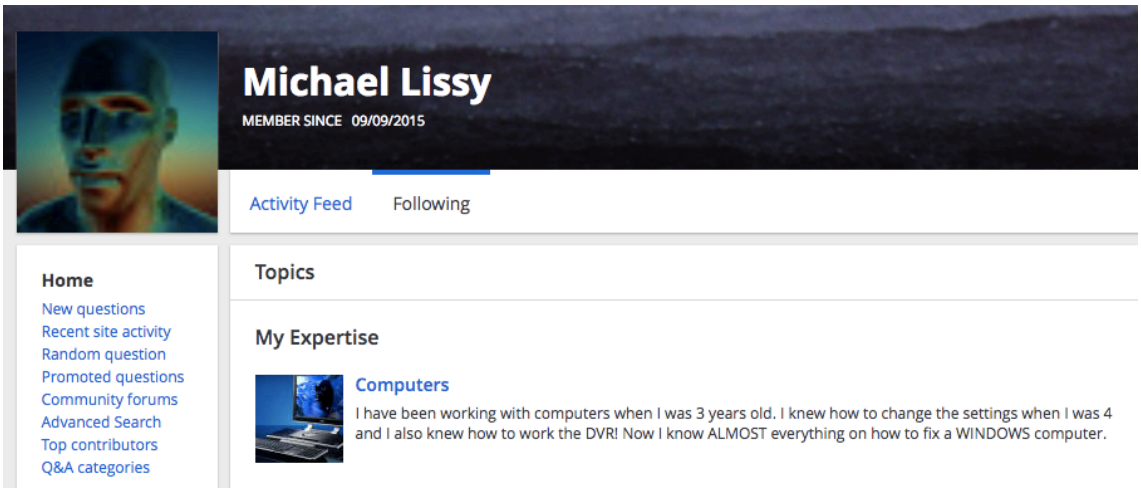

図 5.3 WikiAnswers 上で実際の専門家ユーザと定義したユーザの例の画面キャプチャ[14]

14 https://wiki.answers.com/Q/User:Michael.Lissy.gp8920

5.1.2.1 一般ユーザ

一般ユーザは 5.1.2.1 で示した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

まず, 表 5.1 にて実際に収集した, 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について述べる.

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

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.

以上のことから, 全ての実験において使用する各データ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して評価を行いその平均値をとる. 全ての回答を使用しない理由は, 4.2.1 で説明した通り, 本実験ではコーパスにも Yahoo!知恵袋の質問ページを用いるため, 収集した全ての回答データを評価実験の推定対象にすることができないからである.

5.2 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.

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

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.

5.3 QA サイト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. その後専門性スコアでランキングを生成し, precision@k で評価する. なお, 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, 式(2.1.2)で述べた 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.

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

第6章　　実験結果

第6章では実験結果について述べる. 6.1 にて特定分野を医療とした時の結果について述べる. 6.2 にて特定分野をコンピュータとした時の結果について述べる.6.3 にて考察する.

6.1　特定分野を医療(言語:　日本語)とした時の結果

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. 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 140 まで 10 文字ずつ変化させ, 実験を行なった. 評価は precision@5, precision@10, precision@20 の3種類を用いた. それぞれ結果を図 6.1, 図 6.2, 図 6.3 に示す. また, 手法ごとの推定結果の最大値と最大値の時の使用文字数と推定結果の平均値を表 6.1, 表 6.2, 表 6.3 にまとめる. なお, 対象は長さが 140 文字以上の回答とした.

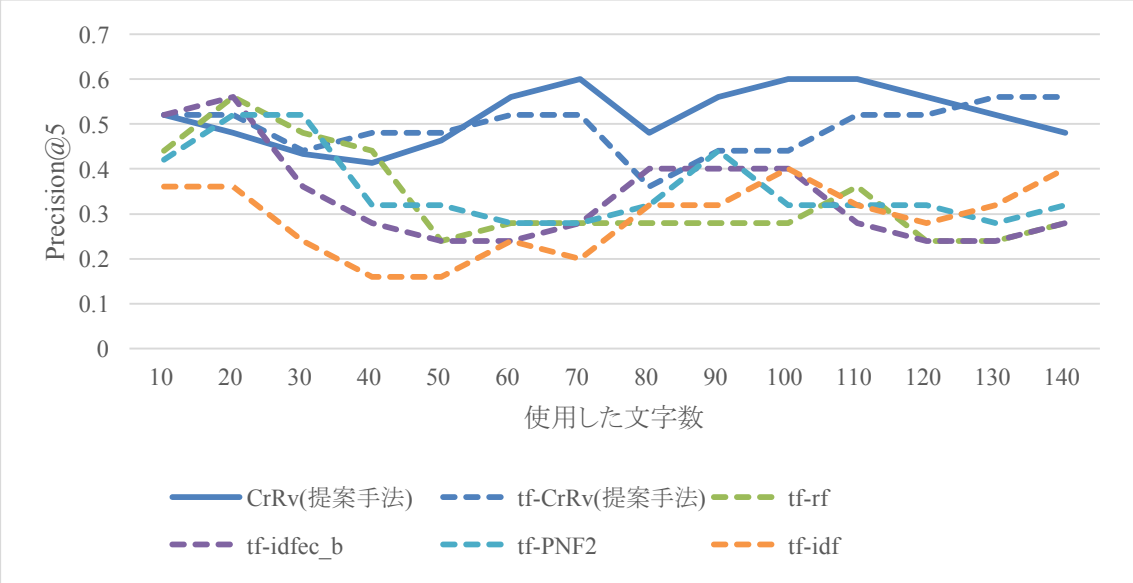

図 6.1 特定分野を医療(言語:　日本語)とした時の, それぞれの文字数を使用した際の precision@5

表 6.1 特定分野を医療(言語:　日本語)とした時の, precision@5 の最大値とその時使用した文字数と平均値

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(precision@5)	**0.60**	0.56	0.56	0.56	0.52	0.40
precision@5 が最大になった時の使用文字数	70	130	20	20	20	100
平均値(precision@5)	**0.52**	0.49	0.33	0.34	0.36	0.29

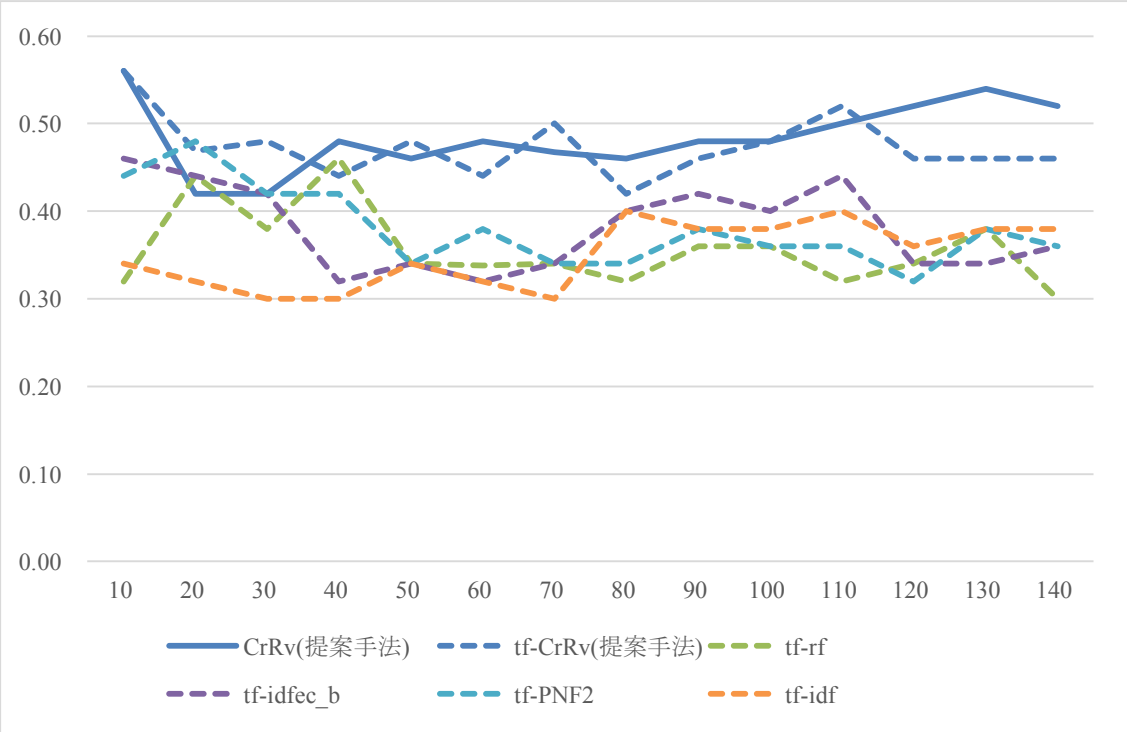

図 6.2 特定分野を医療(言語: 日本語)とした時の, それぞれの文字数を使用した際の precision@10

表 6.2 特定分野を医療(言語: 日本語)とした時の, precision@10 の最大値とその時使用した文字数と平均値

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(precision@10)	**0.56**	**0.56**	0.46	0.46	0.48	0.40

precision@10 が最大になった時の使用文字数	10	10	40	10	20	80
平均値 (precision@10)	**0.48**	0.47	0.36	0.38	0.38	0.35

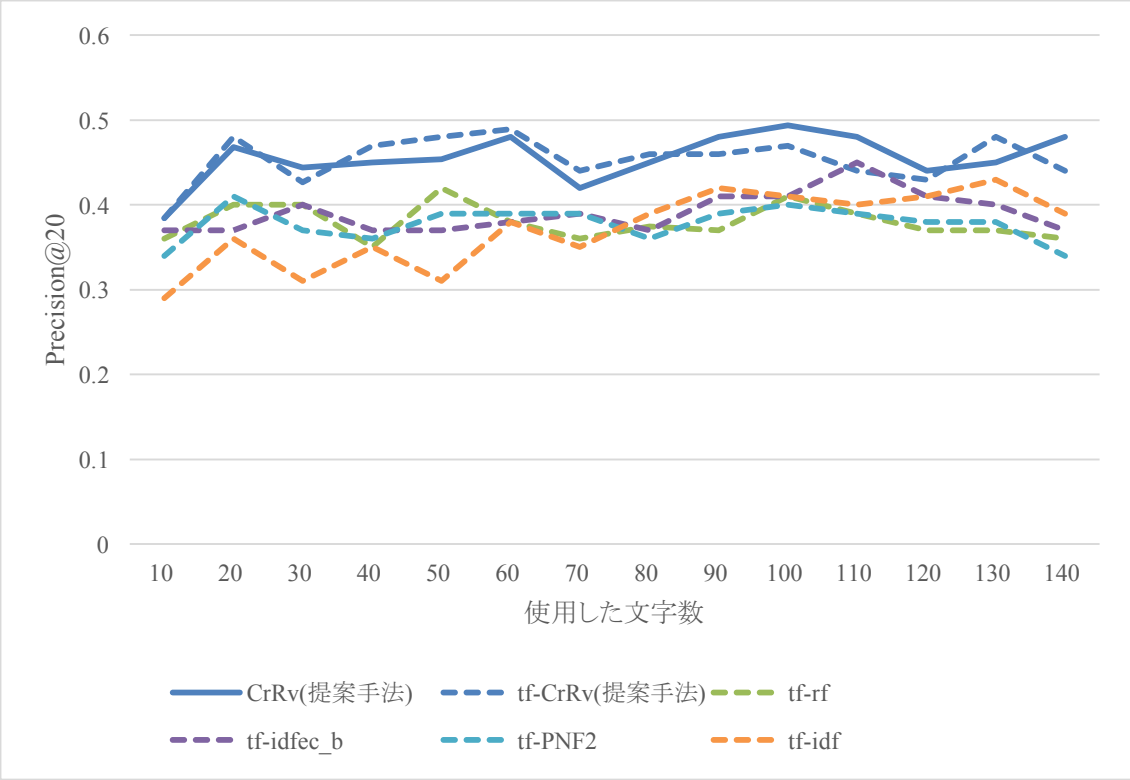

図 6.3 特定分野を医療(言語: 日本語)とした時の，それぞれの文字数を使用した際の precision@20

表 6.3 特定分野を医療(言語: 日本語)とした時の，precision@20 の最大値とその時使用した文字数と平均値

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 (precision@20)	**0.49**	**0.49**	0.42	0.45	0.41	0.43
precision@20 が最大になった時の使用文字数	100	60	50	110	20	130
平均値 (precision@20)	**0.46**	0.45	0.38	0.39	0.38	0.37

ここでデータセットを Yahoo!知恵袋とし，特定分野を医療としたときの結果について考察する．表 6.1～表 6.3 と図 6.1～図 6.3 から全体的に CrRv および tf-CrRv の精度が高い結果となった．提案手法を用いたときの成功例をあげると

> 機能性ディスペプシアとは，胃の痛みや胃もたれなどのつらい症状が続いているにもかかわらず，内視鏡検査などを行っても異常がみつからない病気だそうです．タケキャブは H.ピロリ除菌（1 週間）胃潰瘍，十二指腸潰瘍（6 週間まで）逆理由性食道炎 （8 週間まで）難治性逆流性食道炎（これに限らない）なので，あなたの診断名ではこのお薬では効果はないように思いますし，使えません．機能性デスペプシア でしたら，アコファイド錠が処方される事が多いです

というものがある．これは「機能性ディスペプシア」や「タケキャブ」など知名度の低い単語が適切に重要度が付与され，高い専門スコアを付与することができたと考えられる．一方，一般ユーザの回答に高い専門スコアが付与された例の中には

> セレコックスは，ロキソニンと同じ非ステロイド性鎮痛抗炎症剤に属しますが，ロキソニンより圧倒的に胃腸障害が起きにくい『COX2 選択的阻害剤』というカテゴリに位置づけられ，世界で最も汎用されている非ステロイド性鎮痛抗炎症剤のようです．

というものがある．「阻害剤」，「セレコックス」，「抗炎症剤」などに高い重要度が付与されていた．このユーザはプロフィールをみると会社員としか記載がないため一般ユーザと判断したが，潜在的に専門性が高い可能性がある．

6.2 特定分野をプログラミング(言語: 日本語)とした時の結果

専門家の回答を 500 件，一般人の回答を 2,000 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，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．比率を 1:4 としたのは，収集したデータ全体の数が，専門家の回答が 1,302，一般の回答が 4,093 と約 1:4 となっているからである．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 140 まで 10 文字ずつ変化させ，実験を行なった．評価は precision@5，precision@10，precision@20 の3種類を用いた．それぞれ結果を図 6.4，図 6.5，図 6.6 に示す．また，手法ごとの推定結果の最大値と最大値の時の使用文字数と推定結果の平均値を表 6.4，表 6.5，表 6.6 にまとめる．なお，対象は長さが 140 文字以上の回答とした．

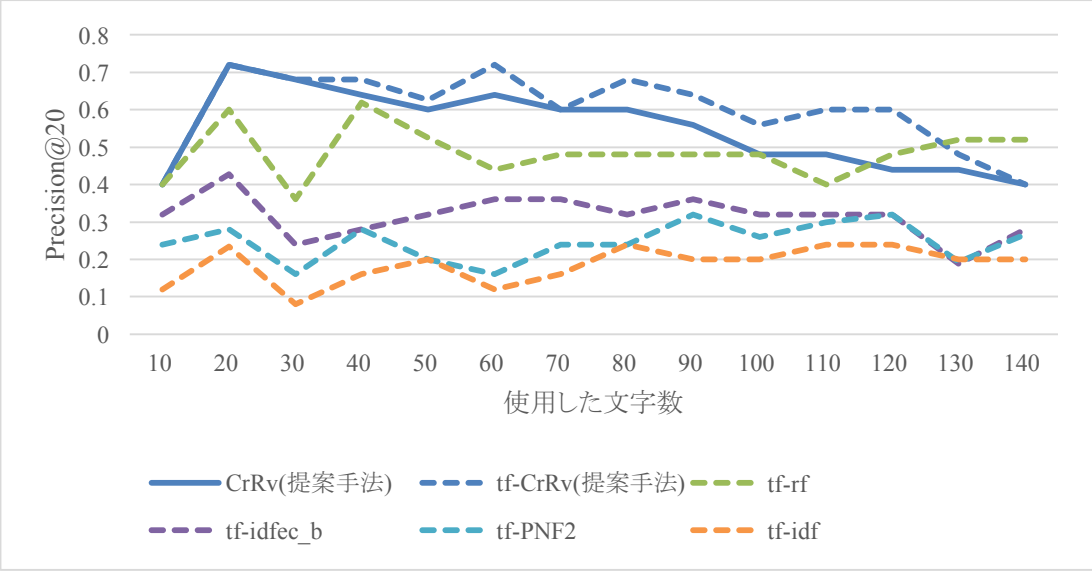

図 6.4 特定分野をプログラミング(言語: 日本語)とした時の，それぞれの文字数を使用した際の precision@5

表 6.4 特定分野をプログラミング(言語: 日本語)とした時の，precision@5 の最大値とその時使用した文字数と平均値

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 (precision@5)	**0.72**	**0.72**	0.62	0.43	0.32	0.24
precision@5 が最大になった時の使用文字数	20	20	40	20	90	80
平均値 (precision@5)	0.55	**0.60**	0.48	0.32	0.25	0.19

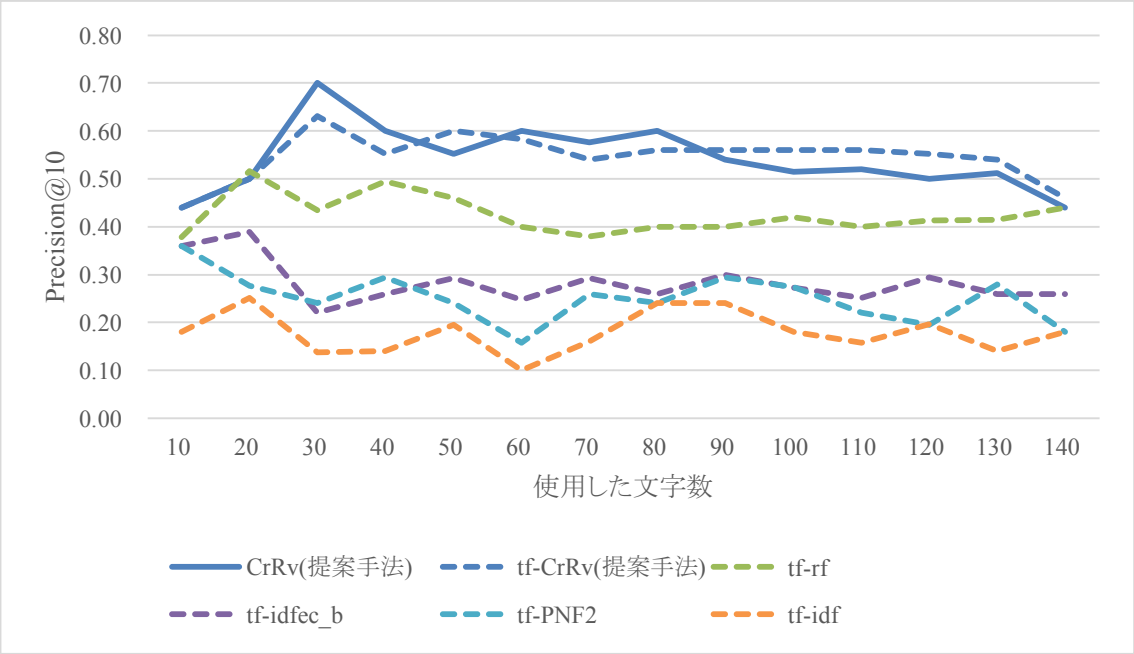

図 6.5 特定分野をプログラミング(言語: 日本語)とした時の，それぞれの文字数を使用した際の precision@10

表 6.5 特定分野をプログラミング(言語: 日本語)とした時の，precision@10 の最大値とその時使用した文字数と平均値

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 (precision@10)	**0.70**	0.63	0.52	0.39	0.36	0.25
precision@10 が最大になった時の使用文字数	30	30	20	20	10	20
平均値 (precision@10)	0.54	**0.55**	0.43	0.28	0.25	0.18

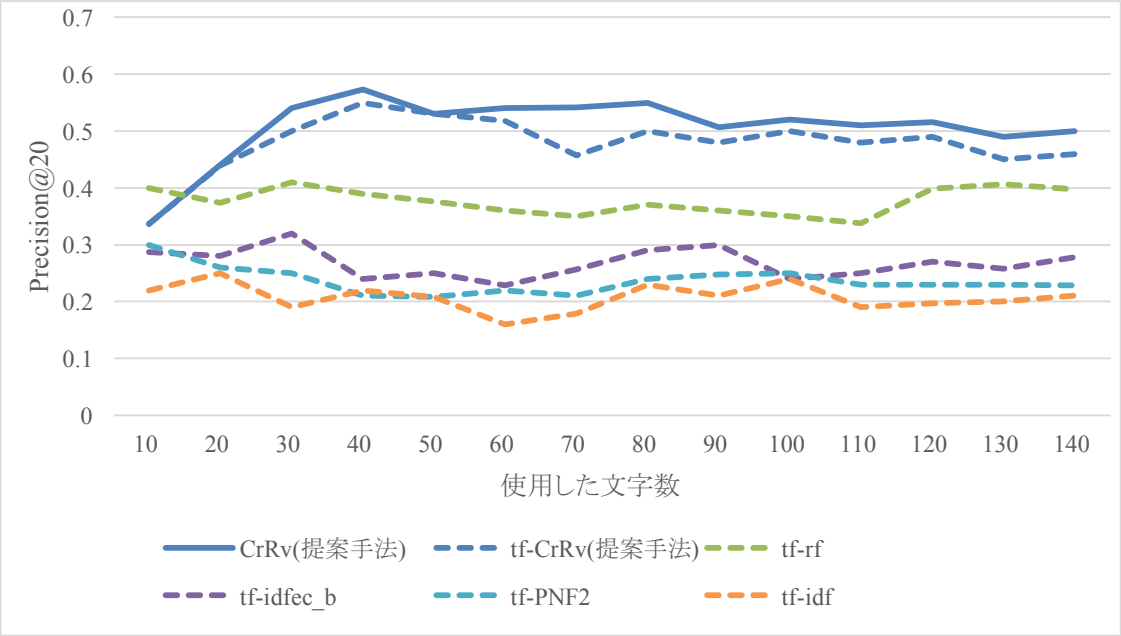

図 6.6 特定分野をプログラミング(言語: 日本語)とした時の，それぞれの文字数を使用した際の precision@20

表 6.6 特定分野をプログラミング(言語: 日本語)とした時の，precision@20 の最大値とその時使用した文字数と平均値

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(precision@20)	**0.57**	0.55	0.41	0.32	0.30	0.25
precision@20 が最大になった時の使用文字数	30	40	30	30	10	20
平均値(precision@20)	**0.51**	0.48	0.38	0.27	0.24	0.21

ここでデータセットを Yahoo!知恵袋とし，特定分野をプログラミングとしたときの結果について考察する．表 6.4～表 6.6 と図 6.4～図 6.6 から全体的に CrRv および tf-CrRv の精度が高い結果となった．CrRv を用いた際の成功例として

PostgreSQL の仕様ではありません．標準 SQL の仕様です．Oracle,SQL Server,MySQL,postgresql すべて同様で，文字列を括るのにダブルクォーテーションは使用できません．″で括るのは，RDBS ごと に意味合いが異なります．

というものがある．これは，「MySQL」「PostgreSQL」「SQL」などに重みが付与されていた．

一方，失敗例としては

レイヤの違いがしっかり理解できていない感じですね・宛先 IP アドレスを 255.255.255.255 や xxx.xxx.xxx.255 等にしてパケットを送信 →IP(TCP/IP)レベルのブロードキャスト・宛先 MAC アドレスを FF:FF:FF:FF:FF:FF にしてパケットを送信→ Ethernet でのブロードキャストです

というものがある．提案手法 CrRv からは「IP」「レイヤ」「アドレス」などに大きな重みが付与されていた．回答ユーザのプロフィールから専門的職業と明確に定義できなかったため一般ユーザとしたが，潜在的に専門性が高いユーザであった可能性がある．

次に，実験データセットを Yahoo!知恵袋とした時について考察する．表 6.1～表 6.6 と図 6.1～図 6.6 から全体的に 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．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．例えば，医療分野の専門家ユーザの回答の中には

アレルゲン(アレルギーの原因物資)となる花粉の種類が多いと，症状が治まるまで，かなりかかると思います．アレルゲンの因子がある以上，症状が出る可能性も高くなりますので，薬は継続して服用してください．また，食物アレルギーでない場合，食生活で改善することはありませんので，気にする必要はありません．私もアレルギー性鼻炎や喘息を患っていますが，処方薬以上に，安全で効果的な対処法はありません．お大事になさってください．

というものがある．含まれる専門用語のみをみると知名度の低い単語は存在しない．したがって，このような回答の専門スコアは低くなってしまう．

6.3 特定分野をプログラミング(言語:英語)とした時の結果

使用するデータセットを WikiAnswers とし，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 6.1，6.2 に使用したデータセットと同じ値にするためである．また，6.3 においては，収集したデータ集合がデータセットを排他的に分けるほど十分なデータが集まらなかったためデータセットは 1 つで行なった．対象とする回答は，カテゴリが「Technology」に属する質問に対しての回答である．使用文字数(character 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@5，precision@10，precision@20 の3種類を用いた．それぞれ結果を図 6.7，図 6.8，図 6.9 に示す．また，手法ごとの推定結果の最大値と最大値の時の

使用文字数と推定結果の平均値を表 6.7，表 6.8，表 6.9 にまとめる．なお，対象は長さが 140 文字以上の回答とした．

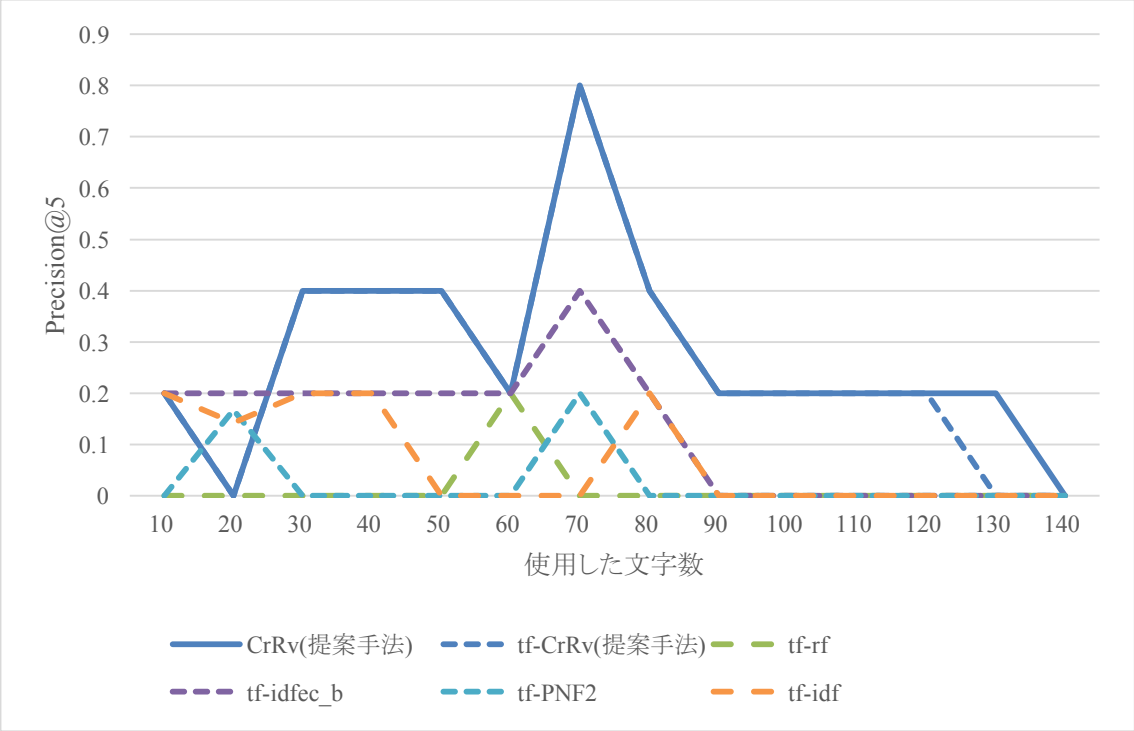

図 6.7 特定分野をプログラミング(言語：英語)とした時の，それぞれの文字数を使用した際の precision@5

表 6.7 特定分野をプログラミング(言語：英語)とした時の，precision@5 の最大値とその時使用した文字数と平均値

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(precision@5)	**0.80**	**0.80**	0.20	0.40	0.20	0.20
precision@5 が最大になった時の使用文字数	70	70	60	70	70	10
平均値(precision@5)	**0.27**	0.26	0.01	0.13	0.03	0.07

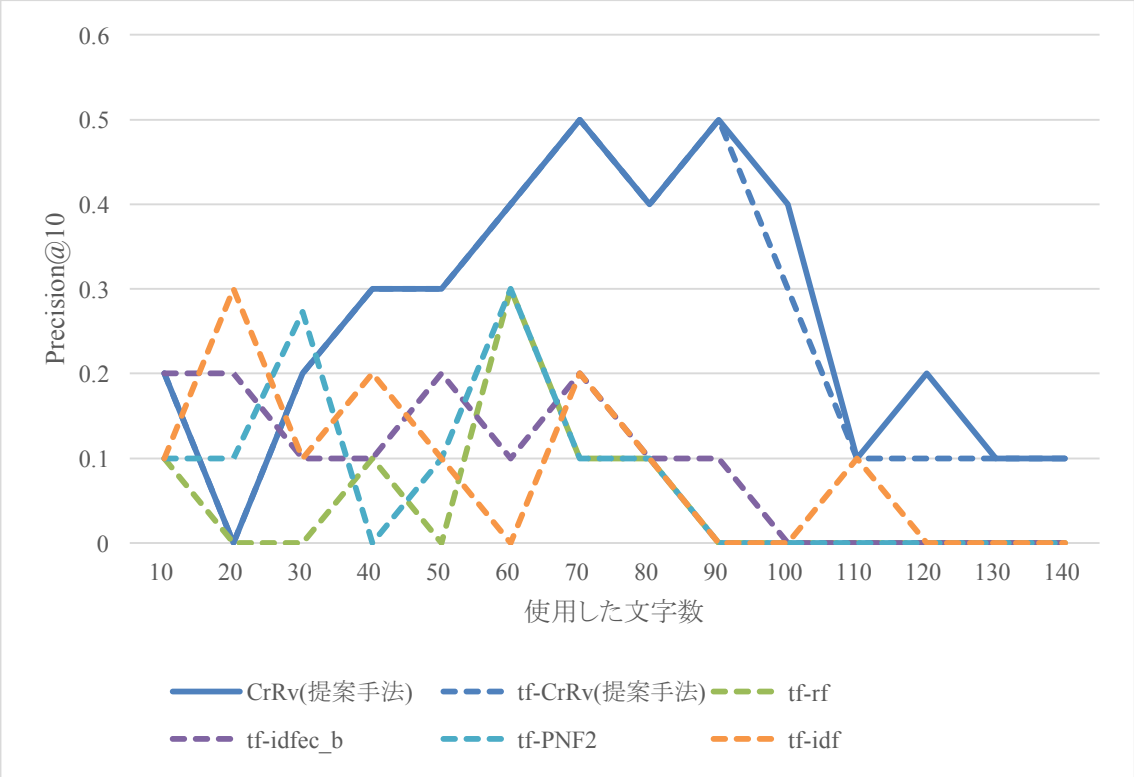

図 6.8 特定分野をプログラミング(言語: 英語)とした時の，それぞれの文字数を使用した際の precision@10

表 6.8 特定分野をプログラミング(言語: 英語)とした時の，precision@10 の最大値とその時使用した文字数と平均値

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 (precision@10)	**0.50**	**0.50**	0.30	0.20	0.30	0.30
precision@10 が最大になった時の使用文字数	70	70	60	10	60	20
平均値 (precision@10)	**0.26**	0.25	0.05	0.09	0.08	0.09

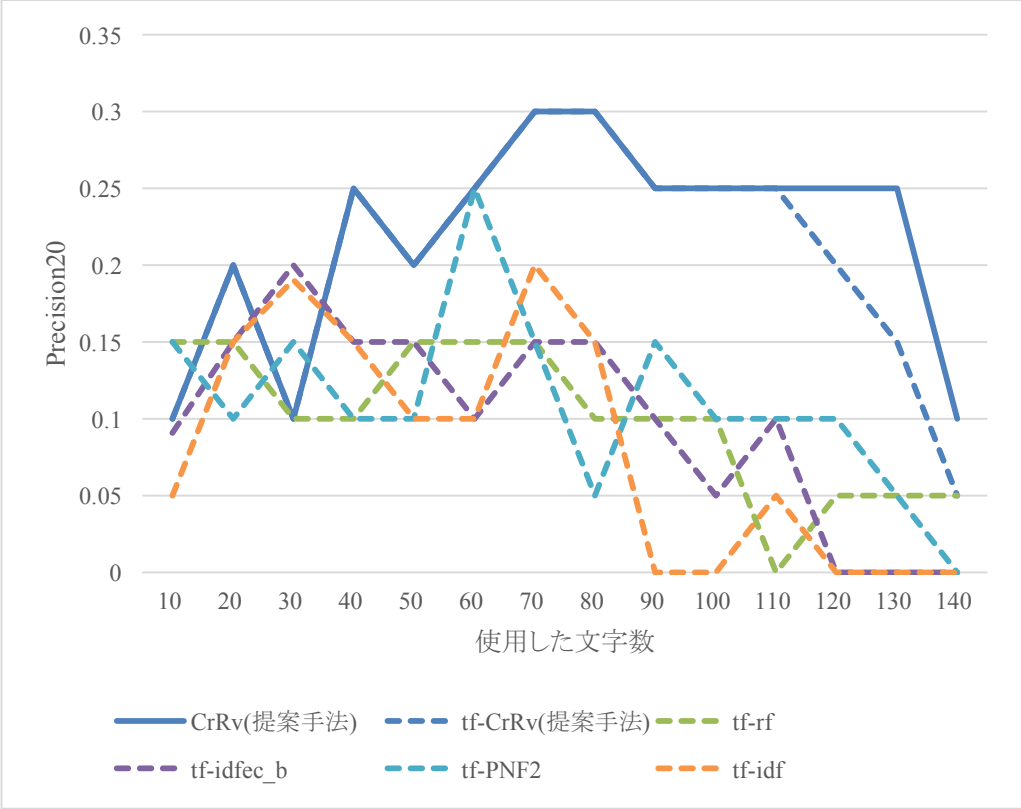

図 6.9 特定分野をプログラミング(言語: 英語)とした時の，それぞれの文字数を使用した際の precision@20

表 6.9 特定分野をプログラミング(言語: 英語)とした時の，precision@20 の最大値とその時使用した文字数と平均値

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 (precision@20)	**0.30**	**0.30**	0.15	0.20	0.25	0.20
precision@20 が最大になった時の使用文字数	70	70	10	30	60	70
平均値 (precision@20)	**0.22**	0.20	0.10	0.10	0.11	0.08

ここで，実験データセットを WikiAnswers とした時について考察する．表 6.7～表 6.9，図 6.7～図 6.9 から全体的に CrRv の精度が高い結果となったことがわかる．しかし，表 6.1～表 6.9，図 6.1～図 6.9 から Yahoo!知恵袋を使用した場合に比べて低いことがわかる．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ため「install」と「installing」は別の単語として判断され，カウントしてしまう．加えて，使用したコーパスの専門性が Yahoo!知恵袋をデー

タセットとした時と大きく異なった可能性がある. Yahoo!知恵袋をデータセットとした場合はコーパスも Yahoo!知恵袋のページを使用して言えるため専門性が同等である. しかし, WikiAnswers をデータセットとした場合は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 を用いた. WikiAnswers と違い StackOverflow はプログラミングに特化した QA サイトである. したがって適応先である WikiAnswers とコーパスとしての専門性が異なっている. さらに, WikiAnswers をデータセットとした時は専門用語を StackOverflow のタグ名と定義した. StackOverflow のタグの中には, 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)も含まれてしまっている. 以上のことから, Yahoo!知恵袋を使用した場合に比べて適切に重みを付与することができなかったと考えられる.

続いて, 成功例と失敗例を示す. 成功例には

Java is an applet program. Netbeans is an IDE, or user interface development program. Jave is part of some of the applications used to make a beans application more enhanced. To get the latest versions packeged together visit sun.java website and look for netbeans version 6 or java and netbeans together.

というものがある. 提案手法 CrRv によって「applet」「netbeans」により大きな重みが付与されていた. 一方, 失敗例には

Escape characters in Java are used in String literals when you need to enter something like a quotation mark, a slash, or a line return. Java escape characters start with the backslash (¥¥). (一部抜粋)

などがある. 「java」「string」「escape」「literal」「characters」といった単語に, それほど大きくはないが重みが付与されていた. 一つ一つは小さい重みだが, 本実験では出現する専門用語の重みを合計する形で専門性スコアを付与するため結果として高い専門性スコアとなってしまったと考えられる.

6.4 考察

本節では全体の考察を行う. 6.4.1 にて使用文字数と精度の関係について考察する. 6.4.2 にて CrRv と tf-CrRv から tf 値について考察する.

6.4.1 使用文字数と精度の関係

使用文字数と精度の関係について考察する. はじめに, 使用した文字数を 30, 50, 70, 90, 110 とした時の評価実験で使用した各手法と各特定分野における precision@10 の結果を表 6.10 でまとめる. なお, 表 6.10 において, 各使用文字数における各特定分野の

表 6.10 使用した文字数を 30, 50, 70, 90, 110 とした時の比較実験で使用した各手法と各特定分野における precision@10 の結果

		文字数 30	文字数 50	文字数 70	文字数 90	文字数 110
CrRv (提案手法)	日本語:医療	0.42	0.46	0.47	**0.48**	0.50
	日本語:プログラミング	**0.70**	0.55	**0.58**	0.54	0.52
	英語:プログラミング	0.20	**0.30**	**0.50**	**0.50**	0.10
tf-CrRv(提案手法)	日本語:医療	**0.48**	**0.48**	**0.50**	0.46	**0.52**
	日本語:プログラミング	0.63	**0.60**	0.54	**0.56**	**0.56**
	英語:プログラミング	0.20	**0.30**	**0.50**	**0.50**	0.10
既存手法での最大値	日本語:医療	0.42	0.34	0.34	0.42	0.44
	日本語:プログラミング	0.43	0.46	0.38	0.40	0.40
	英語:プログラミング	**0.27**	0.20	0.20	0.10	0.10
tf-tf	日本語:医療	0.38	0.34	0.34	0.36	0.32
	日本語:プログラミング	0.43	0.46	0.38	0.40	0.40
	英語:プログラミング	0.00	0.00	0.10	0.00	0.00
tf-idfec_b	日本語:医療	0.42	0.34	0.34	0.42	0.44
	日本語:プログラミング	0.22	0.29	0.29	0.30	0.25
	英語:プログラミング	0.10	0.20	0.20	0.10	0.00
tf-PNF2	日本語:医療	0.42	0.34	0.34	0.38	0.36
	日本語:プログラミング	0.24	0.24	0.26	0.29	0.22
	英語:プログラミング	**0.27**	0.10	0.10	0.00	0.00
tf-idf	日本語:医療	0.42	0.34	0.30	0.38	0.40
	日本語:プログラミング	0.14	0.19	0.16	0.24	0.16
	英語:プログラミング	0.10	0.10	0.20	0.00	0.10

表 6.1～表 6.10, 図 6.1～図 6.9 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, データセットを Yahoo!知恵袋, 特定分野をプログラミングした時は, 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.

続いて表 6.10 について考察する. 表 6.10 から, 多くの場合で既存手法の最大値より提案手法の精度が高いことがわかる. 特に, 使用した文字数が 50 以上の時はすべての場合において既存手法の最大値より提案手法の精度が高い. 例えば, 使用した文字数が 70 の時を見ると, 特定分野を医療(言語: 日本語)とした時は提案手法 CrRv は既存手法の最大値より 0.13, 特定分野をプログ

ラミング(言語: 日本語)とした時は提案手法 CrRv は既存手法の最大値より 0.2, それぞれ precision@10 の値の絶対値が高い. また, 特定分野をプログラミング(言語: 英語)とした時は提案手法 CrRv は既存手法の最大値より 0.3, precision@10 の値の絶対値が高い.

6.4.2 tf 値の有用性

tf 値の有用性について考察する. 表 6.1～表 6.9, 図 6.1～図 6.9 から提案手法の CrRv と tf-CrRv の精度について大きく差がないことがわかる. 今回の対象は短文のため一つの単語が複数出現することが少ない. そのため tf 値を使用しても大きく差が出なかったと考えられる.

第7章　　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場合と，データセットを WikiAnswers，対象特定分野をコンピュータとした場合それぞれにおいて，回答者の専門性の推定を行った．precision@10 で評価を行い，既存手法である tf-rf, tf-PNF2, tf-idfec_b と比較実験を行なったところ，使用文字数を 70 とした際，Yahoo!知恵袋のデータを用いて特定分野を「医療」とした場合で 0.13，特定分野を「プログラミング」とした場合で 0.2 の向上を確認した．また，WikiAnswers のデータを用いて特定分野を「プログラミング」とした場合で 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．

謝辞

本研究を進めるにあたり, 数々のご指導を頂いた山名早人教授に深く厚く本当に御礼申し上げます. また, 研究や実装への助言, 同輩, 石巻優さん, JungKyu Hun 先輩に深く感謝致します.

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).

研究業績

【主著】

- 国内フォーラム（査読あり）

① 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)

FIT 奨励賞およびヤングリサーチャー賞受賞

② 滝川 真弘, 山名 早人 “ノイズに頑健な分野別単語排他度の提案と Twitter ユーザの専門性推定への適用”, データ工学と情報マネジメントに関するフォーラム（DEIM2017）, D5-3（2017.3）

③ 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文からの著者専門性推定への適応”, データ工学と情報マネジメントに関するフォーラム（DEIM2018）, (2018.3) (発表予定)

- 国内ワークショップ（査読なし）

① 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” 研究報告自然言語処理（NL）,2017-NL-233(15),1-6

付録 A: 提案手法 CrRv に用いるパラメータ α，β の試したパターン

第3章で説明した式(3.1.2)内のパラメータ α と式(3.1.6)内のパラメータ β について，試したパターンについて表 A-1 にしめす.

表 A-1 提案手法 CrRv に用いるパラメータ α，β の試したパターン

α	β
0	0
0	1
1	0
1	1
0	$\frac{\sum_{dp}^{DP}\sum_{t'}^{T} tf(t',dp)/\lvert Dp\rvert}{\sum_{dn}^{DN}\sum_{t'}^{T} tf(t',dn)/\lvert Dn\rvert}$
1	$\frac{\sum_{dp}^{DP}\sum_{t'}^{T} tf(t',dp)/\lvert Dp\rvert}{\sum_{dn}^{DN}\sum_{t'}^{T} tf(t',dn)/\lvert Dn\rvert}$
$\frac{\sum_{t'}^{T} n_{Dp}(t')/\lvert Dp\rvert}{\sum_{t'}^{T} n_{Dn}(t')/\lvert Dn\rvert}$	0
$\frac{\sum_{t'}^{T} n_{Dp}(t')/\lvert Dp\rvert}{\sum_{t'}^{T} n_{Dn}(t')/\lvert Dn\rvert}$	1
$\frac{\sum_{t'}^{T} n_{Dp}(t')/\lvert Dp\rvert}{\sum_{t'}^{T} n_{Dn}(t')/\lvert Dn\rvert}$	$\frac{\sum_{dp}^{DP}\sum_{t'}^{T} tf(t',dp)/\lvert Dp\rvert}{\sum_{dn}^{DN}\sum_{t'}^{T} tf(t',dn)/\lvert Dn\rvert}$

付録 B: 実験結果の詳細データ

第6章で示した実験結果の詳細データを示す．なお，表 B-1 から表 B-6 のうち，太文字でかつ下線が引いている値は $p<0.05$ で有意差があるものである．

表 B-1 図 6.1 のグラフにおける詳細データ

使用文字数	CrRv(提案手法)	tf-CrRv(提案手法)	tf-rf	tf-idfec_b	tf-PNF2	tf-idf
10	0.52	0.52	0.44	0.52	0.42	0.36
20	0.48	0.52	0.56	0.56	0.52	0.36
30	0.43	0.44	0.48	0.36	0.52	0.24
40	0.41	0.48	0.44	0.28	0.32	0.16
50	0.46	0.48	0.24	0.24	0.32	0.16
60	0.56	0.52	0.28	0.24	0.28	0.24
70	0.60	0.52	0.28	0.28	0.28	0.20
80	0.48	0.36	0.28	0.40	0.32	0.32
90	0.56	0.44	0.28	0.40	0.44	0.32
100	0.60	0.44	0.28	0.40	0.32	0.40
110	0.60	0.52	0.36	0.28	0.32	0.32
120	0.56	0.52	0.24	0.24	0.32	0.28
130	0.52	0.56	0.24	0.24	0.28	0.32
140	0.48	0.56	0.28	0.28	0.32	0.40

表 B-2 図 6.2 のグラフにおける詳細データ

使用文字数	CrRv(提案手法)	tf-CrRv(提案手法)	tf-rf	tf-idfec_b	tf-PNF2	tf-idf
10	0.56	0.56	0.32	0.46	0.44	0.34
20	0.42	0.47	0.44	0.44	0.48	0.32
30	0.42	0.48	0.38	0.42	0.42	0.30
40	0.48	0.44	0.46	0.32	0.42	0.30
50	0.46	0.48	0.34	0.34	0.34	0.34
60	0.48	0.44	0.34	0.32	0.38	0.32
70	0.47	0.50	0.34	0.34	0.34	0.30

80	0.46	0.42	0.32	0.40	0.34	0.40
90	0.48	0.46	0.36	0.42	0.38	0.38
100	0.48	0.48	0.36	0.40	0.36	0.38
110	0.50	0.52	0.32	0.44	0.36	0.40
120	0.52	0.46	0.34	0.34	0.32	0.36
130	0.54	0.46	0.38	0.34	0.38	0.38
140	0.52	0.46	0.30	0.36	0.36	0.38

表 B-3 図 6.3 のグラフにおける詳細データ

使用文字数	CrRv(提案手法)	tf-CrRv(提案手法)	tf-rf	tf-idfec_b	tf-PNF2	tf-idf
10	0.38	0.38	0.36	0.37	0.34	0.29
20	0.47	0.48	0.40	0.37	0.41	0.36
30	0.44	0.43	0.40	0.40	0.37	0.31
40	0.45	0.47	0.35	0.37	0.36	0.35
50	0.45	0.48	0.42	0.37	0.39	0.31
60	0.48	0.49	0.38	0.38	0.39	0.38
70	0.42	0.44	0.36	0.39	0.39	0.35
80	0.45	0.46	0.37	0.37	0.36	0.39
90	0.48	0.46	0.37	0.41	0.39	0.42
100	0.49	0.47	0.41	0.41	0.40	0.41
110	0.48	0.44	0.39	0.45	0.39	0.40
120	0.44	0.43	0.37	0.41	0.38	0.41
130	0.45	0.48	0.37	0.40	0.38	0.43
140	0.48	0.44	0.36	0.37	0.34	0.39

表 B-4 図 6.4 のグラフにおける詳細データ

使用文字数	CrRv(提案手法)	tf-CrRv(提案手法)	tf-rf	tf-idfec_b	tf-PNF2	tf-idf
10	0.40	0.40	0.40	0.32	0.24	0.12
20	0.72	0.72	0.60	0.43	0.28	0.23
30	**0.68**	0.68	0.36	0.24	0.16	0.08

40	0.64	0.68	0.62	0.28	0.28	0.16
50	0.60	0.63	0.53	0.32	0.20	0.20
60	0.64	0.72	0.44	0.36	0.16	0.12
70	0.60	0.60	0.48	0.36	0.24	0.16
80	0.60	0.68	0.48	0.32	0.24	0.24
90	0.56	0.64	0.48	0.36	0.32	0.20
100	0.48	0.56	0.48	0.32	0.26	0.20
110	0.48	0.60	0.40	0.32	0.30	0.24
120	0.44	0.60	0.48	0.32	0.32	0.24
130	0.44	0.48	0.52	0.19	0.19	0.20
140	0.40	0.40	0.52	0.28	0.27	0.20

表 B-5 図 6.5 のグラフにおける詳細データ

使用文字数	CrRv(提案手法)	tf-CrRv(提案手法)	tf-rf	tf-idfec_b	tf-PNF2	tf-idf
10	0.44	0.44	0.38	0.36	0.36	0.18
20	0.50	0.50	0.52	0.39	0.28	0.25
30	**0.70**	0.63	0.43	0.22	0.24	0.14
40	0.60	0.55	0.49	0.26	0.29	0.14
50	0.55	0.60	0.46	0.29	0.24	0.19
60	0.60	0.58	0.40	0.25	0.16	0.10
70	**0.58**	0.54	0.38	0.29	0.26	0.16
80	**0.60**	0.56	0.40	0.26	0.24	0.24
90	0.54	0.56	0.40	0.30	0.29	0.24
100	0.51	0.56	0.42	0.27	0.27	0.18
110	0.52	0.56	0.40	0.25	0.22	0.16
120	0.50	0.55	0.41	0.29	0.19	0.20
130	0.51	0.54	0.41	0.26	0.28	0.14
140	0.44	0.46	0.44	0.26	0.18	0.18

表 B-6 図 6.6 のグラフにおける詳細データ

使用文字数	CrRv(提案手法)	tf-CrRv(提案手法)	tf-rf	tf-idfec_b	tf-PNF2	tf-idf
10	0.34	0.34	0.40	0.29	0.30	0.22
20	0.44	0.44	0.37	0.28	0.26	0.25
30	**0.54**	0.50	0.41	0.32	0.25	0.19
40	0.57	0.55	0.39	0.24	0.21	0.22
50	0.53	0.53	0.38	0.25	0.21	0.21
60	0.54	0.52	0.36	0.23	0.22	0.16
70	0.54	0.46	0.35	0.26	0.21	0.18
80	0.55	0.50	0.37	0.29	0.24	0.23
90	0.51	0.48	0.36	0.30	0.25	0.21
100	0.52	0.50	0.35	0.24	0.25	0.24
110	0.51	0.48	0.34	0.25	0.23	0.19
120	0.52	0.49	0.40	0.27	0.23	0.20
130	0.49	0.45	0.41	0.26	0.23	0.20
140	0.50	0.46	0.40	0.28	0.23	0.21

表 B-7 図 6.7 のグラフにおける詳細データ

使用文字数	CrRv(提案手法)	tf-CrRv(提案手法)	tf-rf	tf-idfec_b	tf-PNF2	tf-idf
10	0.20	0.20	0.00	0.20	0.00	0.20
20	0.00	0.00	0.00	0.20	0.17	0.14
30	0.40	0.40	0.00	0.20	0.00	0.20
40	0.40	0.40	0.00	0.20	0.00	0.20
50	0.40	0.40	0.00	0.20	0.00	0.00
60	0.20	0.20	0.20	0.20	0.00	0.00
70	0.80	0.80	0.00	0.40	0.20	0.00
80	0.40	0.40	0.00	0.20	0.00	0.20
90	0.20	0.20	0.00	0.00	0.00	0.00
100	0.20	0.20	0.00	0.00	0.00	0.00
110	0.20	0.20	0.00	0.00	0.00	0.00

120	0.20	0.20	0.00	0.00	0.00	0.00
130	0.20	0.00	0.00	0.00	0.00	0.00
140	0.00	0.00	0.00	0.00	0.00	0.00

表 B-8 図 6.8 のグラフにおける詳細データ

使用文字数	CrRv(提案手法)	tf-CrRv(提案手法)	tf-rf	tf-idfec_b	tf-PNF2	tf-idf
10	0.20	0.20	0.10	0.20	0.10	0.10
20	0.00	0.00	0.00	0.20	0.10	0.30
30	0.20	0.20	0.00	0.10	0.27	0.10
40	0.30	0.30	0.10	0.10	0.00	0.20
50	0.30	0.30	0.00	0.20	0.10	0.10
60	0.40	0.40	0.30	0.10	0.30	0.00
70	0.50	0.50	0.10	0.20	0.10	0.20
80	0.40	0.40	0.10	0.10	0.10	0.10
90	0.50	0.50	0.00	0.10	0.00	0.00
100	0.40	0.30	0.00	0.00	0.00	0.00
110	0.10	0.10	0.00	0.00	0.00	0.10
120	0.20	0.10	0.00	0.00	0.00	0.00
130	0.10	0.10	0.00	0.00	0.00	0.00
140	0.10	0.10	0.00	0.00	0.00	0.00

表 B-9 図 6.9 のグラフにおける詳細データ

使用文字数	CrRv(提案手法)	tf-CrRv(提案手法)	tf-rf	tf-idfec_b	tf-PNF2	tf-idf
10	0.10	0.10	0.15	0.09	0.15	0.05
20	0.20	0.20	0.15	0.15	0.10	0.15
30	0.10	0.10	0.10	0.20	0.15	0.19
40	0.25	0.25	0.10	0.15	0.10	0.15
50	0.20	0.20	0.15	0.15	0.10	0.10
60	0.25	0.25	0.15	0.10	0.25	0.10
70	0.30	0.30	0.15	0.15	0.15	0.20

80	0.30	0.30	0.10	0.15	0.05	0.15
90	0.25	0.25	0.10	0.10	0.15	0.00
100	0.25	0.25	0.10	0.05	0.10	0.00
110	0.25	0.25	0.00	0.10	0.10	0.05
120	0.25	0.20	0.05	0.00	0.10	0.00
130	0.25	0.15	0.05	0.00	0.05	0.00
140	0.10	0.05	0.05	0.00	0.00	0.00