

確率的言語モデルによる意見文抽出のための局所文脈スムージング

本田 徹也 †, 江口浩二 ‡
† 神戸大学自然科学研究科, ‡ 神戸大学工学研究科

要旨

近年、情報技術の進歩によって数多くの文書が電子化され、人々の間で共有されるようになった。このような文書や文を検索する手法の中で重要なものの1つに意見文検索がある。ある着目した話題に対する肯定的もしくは否定的な意見を検索できることは非常に有用である。本研究では、文書中の局所的な範囲において意見極性が変化しないことが多いという特性に着目し、局所文脈という概念を導入し意見文検索に応用する。また、検索有効性を高める上でも用いられる従来の言語モデルの平滑化手法を拡張し、局所文脈の言語モデルも含めた手法を提案する。これにより、従来手法と比較して意見文の検索有効性が向上することを実験によって示す。

Locally contextualized smoothing for sentiment sentence extraction using probabilistic language models

Tetsuya HONDA†, Koji EGUCHI ‡
† Kobe University, Graduate School of Science and Technology
‡ Kobe University, Graduate School of Engineering

Abstract

Recently a number of documents are published on the web. One of crucial techniques to access to the web content is sentiment sentence retrieval. It is very useful to retrieve positive or negative opinions to a specific topic at sentence level. Considering the property that sentiment polarities are often locally consistent in a document, we focus on using a local context information for the sentiment sentence retrieval. For this objective, we propose a new smoothing method, extending Dirichlet smoothing, to improve effectiveness of the retrieval. We show through experiments that the proposed method is more effective than conventional methods for the task of sentiment sentence retrieval.

1 はじめに

近年、情報技術の進歩によって数多くの文書が電子化され、多数の人々の間で共有されるようになった。このような文書や文を検索する手法は多くの研究者達によって考案されている。その中でも特に注目を浴びている検索手法の1つとして、意見文検索がある。意見文検索とは、ある話題に対して肯定的な文や否定的な文を検索することである。この技術を利用することで、Web上に掲載されているある特定の商品についての評判、誹謗中傷情報の検出などについての精度向上が期待されている。

これまでの意見文検索の研究では、意見の話題依存性がほとんど考慮されていなかった [1, 2]。しかし [3] では、話題と意見のクエリ語を別に扱い、それぞれを文検索に用いて検索結果を結合するこ

とで、意見の話題依存性を考慮している。さらに、ある話題に対しての意見の方向性を示す意見極性を指定した意見文検索も提案している。しかし、検索有効性の面からは改善の余地がある。

そこで本研究では、1つの文書中で意見極性が頻繁に変化しないという特性に着目し、局所文脈という概念を導入することで意見文検索の有効性の向上を目指す。また、通常の文レベルの情報検索においても、局所文脈を考慮した平滑化による検索有効性の改善が期待される。提案手法を意見文検索に適用することで、従来手法と比較して性能が向上することを実験によって示す。

2 意見文検索

本章では、意見文検索の定義とその意味について述べる。そのために、まず文検索についての定

義についても触れ、検索モデルについて言及する。また、これらの形式的な定義を与える。

2.1 文検索

文検索は、検索対象の文集合からユーザのもつ情報ニーズに適合する文を検索する処理のことである。検索された文は、情報ニーズと文との適合度によって順序付けられユーザに提示される。

以降では、検索対象の文を確率的言語モデルとして扱ったクエリ尤度モデルと適合モデルについて説明する。

2.1.1 クエリ尤度モデル

確率的言語モデルを用いた検索モデルの1つとして、文モデルからクエリ語を生成する尤度を適合度とする検索モデルがある。このような検索モデルをクエリ尤度モデル [4, 5] という。文モデル S からクエリ Q を生成する尤度 $P(Q|S)$ は式 (1) のように定義される。

$$P(Q|S) = \prod_{q \in Q} P(q|S)^{c(q,Q)} \quad (1)$$

ここで $c(q, Q)$ は、クエリ Q の中でクエリ語 q が出現する回数である。文モデル S からクエリ語 q を生成する尤度 $P(q|S)$ は最尤推定法によって推定する。

式 (1) より、文モデル S の中に全く含まれないクエリ語 q が存在すると、他にどのようなクエリ語が存在しようとも尤度 $P(Q|S)$ が必ず 0 になることに注目してほしい。尤度が 0 となり他の文との比較が困難になる状況をゼロ確率問題という。ゼロ確率問題を回避する手法の1つとして、言語モデルの平滑化がよく行われている。これについては第3章で説明する。

2.1.2 適合モデル

仮に検索対象の文集合を適合クラスと不適合クラスに分割できるとする。このとき、適合モデル [6] は適合クラスに属する文書を用いて推定された、検索要求に関する言語モデルである。これをクエリとして用いて文書集合から適合文書を検索する。現実には適合クラスを完全に把握することは困難なため、ユーザから与えられるクエリから適合文モデルを推定し、検索対象の文書を順序付ける。

文献 [6] に倣って、以下に検索処理の単位を文書と想定した定義を与える。クエリを $Q = \{q_1, q_2, \dots, q_n\}$ 、適合クラスを R とするとき、適

合モデルからある語 w が選択される確率は次のようになる。

$$\begin{aligned} P(w|R) &= P(w|q_1, q_2, \dots, q_n) \quad (2) \\ &= \frac{P(w, q_1, q_2, \dots, q_n)}{P(q_1, q_2, \dots, q_n)} \quad (3) \end{aligned}$$

ここで、クエリ q_1, q_2, \dots, q_n と同時に語 w を選択する確率を推定するために、クエリと語はそれぞれ独立に適合モデルから選択されると仮定する。

文書集合 C からある文書 D を選択する確率を $p(D)$ とするとき、文書 D から $n+1$ 回語を選択する場合を考える。クエリ q_1, q_2, \dots, q_n と共に語 w を選択する確率を式 (4) に示す。

$$P(w, q_1, q_2, \dots, q_n) = \sum_{D \in C} p(D) P(w, q_1, q_2, \dots, q_n | D) \quad (4)$$

クエリ q_1, \dots, q_n と語 w はそれぞれ独立に選択しているため、式 (4) の $P(w, q_1, \dots, q_n | D)$ は次のようにも表わすことができる。

$$P(w, q_1, q_2, \dots, q_n | D) = P(w | D) \prod_{i=1}^k P(q_i | D) \quad (5)$$

式 (5) を式 (4) に代入すると、次のように表わせる。

$$P(w, q_1, q_2, \dots, q_n) = \sum_{D \in C} p(D) P(w | D) \prod_{i=1}^k P(q_i | D) \quad (6)$$

これにより、式 (6) を式 (3) に代入することで、次の式を導ける。

$$P(w|R) = \frac{\sum_{D \in C} p(D) P(w | D) \prod_{i=1}^k P(q_i | D)}{P(q_1, \dots, q_n)} \quad (7)$$

なお、ベイズの定理を用いると、語 w を適合モデルから観測する確率 $P(w|R)$ は次の式のように表わせる。

$$P(w|R) = \sum_{D \in C} P(w | D) P(D | q_1, \dots, q_n) \quad (8)$$

現実には適合モデルを用いる場合は、 $P(w|R)$ は次のような近似によって求められることが多い。 $P(D | q_1, \dots, q_n)$ を用いて降順に順序付けされた文書の上位 N 個を利用して、式 (8) により混合分布を構築する。構築された混合分布から、 $P(w|R)$ の値が高い語 w のうち、 M 語を用いる。

さて、適合モデルを文検索に応用することを考える。これまでの議論における文書 D を文 S に置き換えることで容易に文に基づく適合モデルが推定できる。

適合モデルを用いて検索対象の文集合を順序付けるには、文モデルと適合モデルとの非類似度を示すクロスエントロピーを用いる。

文書集合に含まれる語彙集合を ν とした時、適合クラスの言語モデル R と文モデル S とのクロスエントロピー $H(R||S)$ は次のように定義される。

$$H(R||S) = - \sum_{w \in \nu} P(w|R) \log P(w|S) \quad (9)$$

式 (9) から、 $-H(R||S)$ の降順に文を順序付ける。適合モデルはクエリ尤度モデルと比較して精度が良いことが知られている [6]。

2.2 意見文検索

以降では、第 2.1.1 節、第 2.1.2 節で紹介したクエリ尤度モデルと適合モデルを、意見文検索のために拡張したクエリ尤度モデルと意見適合モデルについて文献 [3] に基づいて概説する。

2.2.1 意見クエリ尤度モデル

ある確率的言語モデル $\mathbf{p}_1, \mathbf{p}_2$ と意見極性モデル \mathbf{p}_x に対してある確率を与える関数 $\pi(\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_x)$ を定義する。ある文 $S_i = \{w_{i1}, w_{i2}, \dots, w_{im_i}\}$ に含まれる語 w_{ij} が話題を示す語か意見を示す語かを示す変数 $b_{ij} \in \{S, T\}$ を導入する。意見極性分布から、ある意見極性 x_i を観測する確率を $\mathbf{p}_x(x_i)$ とするとき、新たな文 $S_i = \{w_{i1}, w_{i2}, \dots, w_{im}\}$ を観測する確率は次のように表わせる。

$$\sum_{\mathbf{p}_t, \mathbf{p}_s, \mathbf{p}_x} \pi(\mathbf{p}_t, \mathbf{p}_s, \mathbf{p}_x) \mathbf{p}_x(x_i) \prod_{j=1}^m \begin{cases} \mathbf{p}_t(w_{ij}) & \text{if } b_{ij} = T \\ \mathbf{p}_s(w_{ij}) & \text{otherwise} \end{cases} \quad (10)$$

ただし、 \mathbf{p}_t は話題に関する確率的言語モデル、 \mathbf{p}_s は意見に関する確率的言語モデルである。ここで、以降の説明の便宜上、指示関数 1_y を定義する。指示関数 1_y は、述語 y が真であれば 1、そうでなければ 0 を返す。関数 $\pi(\mathbf{p}_t, \mathbf{p}_s, \mathbf{p}_x)$ を、確率的言語モデル $\mathbf{p}_t, \mathbf{p}_s, \mathbf{p}_x$ が文の確率的言語モデル $\mathbf{p}_{ti}, \mathbf{p}_{si}, \mathbf{p}_{xi}$ に対応するときに $\frac{1}{n}$ を返す関数として定義する。

$$\pi(\mathbf{p}_t, \mathbf{p}_s, \mathbf{p}_x) = \frac{1}{n} \sum_{i=1}^n 1_{\mathbf{p}_t=\mathbf{p}_{ti}} \times 1_{\mathbf{p}_s=\mathbf{p}_{si}} \times 1_{\mathbf{p}_x=\mathbf{p}_{xi}} \quad (11)$$

ここで、 n は文書コレクションに含まれる文の総数である。式 (11) は、話題に関する言語モデル \mathbf{p}_t と意見に関する言語モデル \mathbf{p}_s が共起する場合に 0 より大きい値を割り当てる。そのため、両モデル間の経験的な依存性を表現していると言える。

本研究の以降の実験では意見極性分布を用いない検索モデルを対象にするため、式 (10) の文書モデルがクエリを生成する尤度は次式のように簡約化できる。

$$\prod_{w \in \mathbf{q}^t} \mathbf{p}_t(w)^{c(w, \mathbf{q}^t)} \prod_{w \in \mathbf{q}^s} \mathbf{p}_s(w)^{c(w, \mathbf{q}^s)} \quad (12)$$

ここで、 \mathbf{q}^t はユーザの与えるクエリの中の話題についての語集合、 \mathbf{q}^s はクエリの中の意見についての語集合である。式 (12) の対数を取り、さらに話題部分と意見部分に重みづけできるよう拡張することで次式を得る。

$$\alpha \sum_{w \in \mathbf{q}^t} \mathbf{q}^t(w) \log \mathbf{p}_t(w) + (1-\alpha) \sum_{w \in \mathbf{q}^s} \mathbf{q}^s(w) \log \mathbf{p}_s(w) \quad (13)$$

本論文では、式 (13) の形式を意見クエリ尤度モデルと称する。

2.2.2 意見適合モデル

前節で述べた意見クエリ尤度モデルは、適合モデルの概念を拡張することで意見文検索にも応用できる。以降に形式的な定義を与える。文の集合 C とユーザが与えたクエリ $\{\mathbf{q}^t, \mathbf{q}^s, \mathbf{q}^x\}$ があるとす。ここで、 \mathbf{q}^t は話題を示すクエリ語、 \mathbf{q}^s は意見極性を典型的に表現するシードワード、 \mathbf{q}^x は求める意見極性を二値で表現したものである。ユーザの情報ニーズに含まれる話題に関する適合モデル \mathbf{r}_t と意見に関する適合モデル \mathbf{r}_s は次のように推定する。クエリ $\{\mathbf{q}^t, \mathbf{q}^s, \mathbf{q}^x\}$ を式 (10) で表わされた確率分布から無作為に抽出されたとし、さらにもうひとつの話題に関する語、もしくは意見に関する語 w を抽出するときの尤度は次のようになる。

$$\mathbf{r}_t(w) = \frac{P(\mathbf{q}^t \circ w, \mathbf{q}^s, \mathbf{q}^x)}{P(\mathbf{q}^t, \mathbf{q}^s, \mathbf{q}^x)} \quad (14)$$

$$\mathbf{r}_s(w) = \frac{P(\mathbf{q}^w, \mathbf{q}^s \circ w, \mathbf{q}^x)}{P(\mathbf{q}^t, \mathbf{q}^s, \mathbf{q}^x)} \quad (15)$$

ここで、 $\mathbf{q} \circ w$ は語の列 \mathbf{q} に語 w を追加する操作のことを示す。

話題に関する適合モデル \mathbf{r}_t と意見に関する適合モデル \mathbf{r}_s を推定した後、それらを用いて文を順序付ける。適合モデルではクロスエントロピーを用いて順序付けを行ったが、意見適合モデルではクロスエントロピーを変形して順序付けに応用する。

$$\alpha \sum_{w \in \nu} \mathbf{r}_t(w) \log \mathbf{p}_t(w) + (1-\alpha) \sum_{w \in \nu} \mathbf{r}_s(w) \log \mathbf{p}_s(w) \quad (16)$$

ここで α は話題と意見とのバランスを調整するパラメータ、 ν を文書集合に含まれる語彙集合とする。このパラメータ α は訓練データを用いて経験的に定める。

3 関連研究

文書検索のための言語モデルの平滑化手法の目的は大きく分けて 2 つ存在すると考えられる。

1. クエリを生成する尤度の下限を 0 よりわずかに大きくすることによりゼロ確率問題を回避する。
2. データスパースネスによる文書モデルの推定の難しさを改善する。

平滑化の研究は多くなされており、さまざまな種類の手法が存在する。しかし、着目される点はそのような割合で言語モデルを結合し、混合モデルを生成するかという点であるといえる。

以降では、文書検索のための平滑化手法の中でも特に効果が高いとされる Dirichlet スムージング [7] について説明する。

Dirichlet スムージングは、線形補間法に対して言語モデルの述べ語数も加味した手法である。線形補間法は、補間対象の言語モデルの述べ語数が大きく異なる場合でも同じ比率で補間する。そのため、述べ語数が小さい言語モデルに対しては補間が不足し、逆に述べ語数が大きい言語モデルに対しては補間しすぎる傾向がある。そこで、Dirichlet スムージングでは検索モデルの混合比を表わすスムージングパラメータに改良が施されている。これにより、スムージングパラメータが平滑化に与える影響を適度に軽減し、語数の特徴に差異のある言語モデルに対しても検索有効性の向上が望める平滑化が行える。

以降で形式的な定義を与える。文モデル S から語 w を生成する尤度 $\hat{P}(w|S)$ は、平滑化を行うと式 (17) のように定義される。

$$\begin{aligned} \hat{P}(w|S) &= \lambda P(w|S) + (1-\lambda)P(w|C) \quad (17) \\ \lambda &= \frac{|S|}{|S| + \alpha} \quad (18) \end{aligned}$$

ここで、 $|S|$ は文モデル S に含まれる語数である。また、 α はスムージングパラメータである。スムージングパラメータ α は 0 以上の値をとる。

4 局所文脈を用いたスムージング

4.1 局所文脈とは

書籍などのような形式的に整った文書では、限定された特定の話題について述べられており、その話題についてもある立場に基づいて記述されている場合が多い。しかし、Web 上に公開されたブログのような文書では、形式的にも様々な質の違いがあり、必ずしも特定の話題について言及されているわけではなく、また特定の意見極性を示す意見のみが記述されているわけではない。そのような文書からも有益な情報を持つ文を抽出するためには、より文脈に着目した確率的言語モデルを構築することが重要である。文書中の文脈に着目することで、徐々に変化する話題や立場についても考慮された言語モデルを生成することが可能となる。

そこで本研究では、局所文脈という概念を導入する。局所文脈は、着目する文に対して前後に存在する複数の文も含む文集合のことである。以降に形式的な定義を与える。文書モデルを $D = \{S_1, S_2, \dots, S_n\}$ 、着目する文モデルを S_k とする時、局所文脈モデル $LC_k = LC(k, width, D)$ からクエリ Q を生成する尤度 $P(Q|LC_k)$ を次のように定義する。

$$P(Q|LC_k) = \prod_{q \in Q} P(q|LC_k)^{c(q, LC_k)} \quad (19)$$

$$P(q|LC_k) = \frac{\sum_j |S_j| \hat{P}(q|S_j)}{\sum_i |S_i|} \quad (20)$$

ここで、 $\sum_i = \sum_{i=\max(1, k-width)}^{\min(n, k+width)}$ である。また、 $width$ は局所文脈の幅を示すパラメータであり、非負の値をとる。局所文脈の幅 $width$ を限りなく大きく設定することで、局所文脈モデルは文書モデルと等価となる。

4.2 局所文脈を用いた平滑化

局所文脈を文検索で利用する場合、第 3 章で紹介した平滑化手法では 2 つの言語モデルに対してのみ考慮されているため、そのままでは十分な効果が期待することはできない。そこで、このような平滑化手法を応用し、3 つの言語モデルに対して適用できる平滑化手法を提案する。

提案手法では、これまで検索有効性のよい結果を示した Dirichlet スムージング手法を拡張する。平滑化される文モデルを S 、その文を含む文書を D 、平滑化される文モデルから語 w を生成する尤度を $P(w|S_k)$ 、局所文脈モデルから語 w を生成

する尤度を $P(w|LC_k)$, 文書コレクションモデルから語 w を生成する尤度を $P(w|C)$ とするとき, 平滑化された文モデルから語 w を生成する尤度 $\hat{P}(w|S_k)$ を次のように定義する.

$$\hat{P}(w|S_k) = \lambda P(w|S_k) + (1 - \lambda) \hat{P}(w|LC_k) \quad (21)$$

$$\hat{P}(w|LC_k) = \pi P(w|LC_k) + (1 - \pi) P(w|C) \quad (22)$$

ここで, λ と π は次のように定義する.

$$\lambda = \frac{|S|}{|S| + \alpha} \quad (23)$$

$$\pi = \frac{|LC|}{|LC| + \beta} \quad (24)$$

$|S|$ は文モデル S に含まれる延べ語数であり, $|LC|$ は局所文脈モデル LC に含まれる延べ語数である. また, α, β はスムージングパラメータであり, 非負の値をとる. スムージングパラメータを大きく設定すればするほど平滑化の度合いが少なくなり, 文モデル S の分布が強く現れる言語モデルが生成される. このスムージングパラメータは, 実験により最適な値を求める.

ただし, 局所文脈の幅 *width* が 0 の場合は局所文脈は存在しないものと考え, 次のように定義する.

$$\hat{P}(w|S_k) = \frac{|S|P(w|S_k) + \alpha P(w|C)}{|S| + \alpha} \quad (25)$$

式 (25) は, 文モデルの尤度 $P(w|S_k)$ に対して文書コレクションモデルの尤度 $P(w|C)$ を用いて Dirichlet スムージングした場合と等価である. 第 5 章では, スムージングパラメータの最適値, また局所文脈の最適な幅についても実験を行い評価する.

5 実験と評価

5.1 評価実験

提案手法の有効性を検証するために, いくつかの評価実験を行う. 評価実験を行うにあたり, 次に挙げる 2 つの点について着目する.

- 提案手法で用いる局所文脈において, 精度が最も向上する幅はどの程度か.
- 文検索の応用としての意見検索において, 従来手法と比較してどの程度精度が向上したか.

評価実験では, MPQA Opinion Corpus version 1.2 [8, 9] と呼ばれるニュース記事を用いて, 言語モデルの構築とそのモデルの評価を行った. このテキストデータには, 187 ヶ国の 2001 年 6 月から 2002 年 5 月までの英文ニュースで構成されて

おり, 535 文書, 11,114 文から成る. また, このテキストデータには意見などを示すフレーズにアノテーションが附加されている.

MPQA corpus は, 話題の適合判定については文書単位で附加されている. そこで, 本実験では適合文書に含まれる全ての文を適合と仮定した. また, 意見の適合判定はフレーズ単位で附加されている. そのため, それぞれの文の中に含まれる意見強度の最も強い意見極性を, その文の意見極性と仮定した.

クエリの意見語として与えるシードワードに, KAM [10], TUR [11] および LAV [3] の 3 種類を用いた. それぞれの語集合は, 肯定属性と否定属性を持つ語集合から成る.

さらに, 話題を示すクエリを 10 種類用意した. 話題を示すクエリとシードワードを同時に与える検索タスクの場合, そのクエリに肯定属性を示すシードワードを追加する場合と否定属性を示すシードワードを追加する場合の 2 つの組み合わせについて実験する. つまり, 全 20 種類のクエリを文検索システムに与えて実験を行う.

意見を示すクエリのみを用いる場合は, 肯定属性を示すシードワードを与える場合と, 否定属性を示すシードワードを与える場合の 2 種類について実験を行う.

実験に MPQA corpus を用いるにあたり, 文中に含まれる単語, またクエリに対して Krovetz stemmer [12] を適用し, さらに 418 個のストップワードを除去した.

提案手法がどのような意見文検索タスクについて有効かを比較するために, いくつかの検索タスクを設定する.

lmtf 標準的な言語モデル手法のクエリ尤度モデルの下で, 話題と意見を分割しない検索対象データに対して, 話題のみを示すクエリを用いて検索する方法.

lms クエリ尤度モデルの下で, 意見記述部のみを抽出した検索対象データに対して, 意見のみを示すクエリを用いて検索する方法.

basef lmtf と lms の線形結合.

rmtf 話題と意見を分割しない検索対象データに対して, 話題のみを示すクエリを用いて推定した従来型の適合モデル [6].

rmt-basef 話題と意見を分割しない検索対象データを, basef の検索結果をクエリとして用いた適合モデル.

rms-basef 意見記述部のみを抽出した検索対象

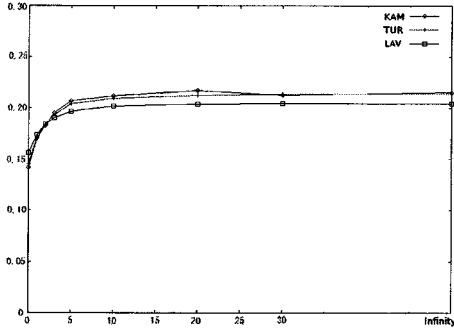


Fig. 1: The result of experiments for basef with Bpref

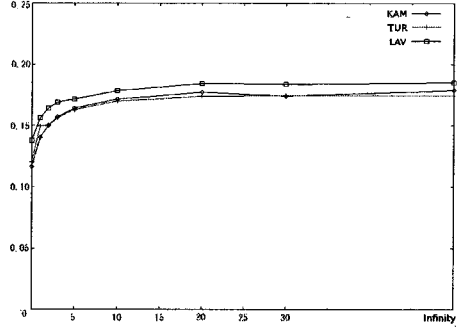


Fig. 2: The result of experiments for basef with MAP

Table 1: The table of experiment results for basef with Bpref/MAP. Left values are results evaluated with Bpref, and right values are results evaluated with MAP.

width	KAM		TUR		LAV	
0	0.1416	(0.1142)	0.1451	(0.1188)	0.1550	(0.1369)
1	0.1702	(0.1404)	0.1710	(0.1415)	0.1740	(0.1562)
2	0.1829	(0.1501)	0.1841	(0.1498)	0.1840	(0.1642)
3	0.1950	(0.1573)	0.1933	(0.1565)	0.1899	(0.1692)
5	0.2067	(0.1643)	0.2041	(0.1630)	0.1967	(0.1715)
10	0.2120	(0.1714)	0.2093	(0.1699)	0.2020	(0.1785)
20	0.2171	(0.1775)	0.2125	(0.1739)	0.2040	(0.1842)
30	0.2126	(0.1740)	0.2137	(0.1746)	0.2046	(0.1839)
Infinity	0.2155	(0.1792)	0.2136	(0.1746)	0.2044	(0.1852)

データに対して、basef の検索結果をクエリとして用いた適合モデル。

slmf rmt-basef と rms-basef の線形結合。

実験では、前述の検索タスクのうち、lms を除いた 6 つについて評価を行う。さらにこれらの意見文検索タスクに対して、2 つの実験フェーズを設定し評価する。

訓練フェーズ 与えられた検索対象データとクエリに対して検索有効性が最良となるように、パラメータを設定する。

テストフェーズ 設定されたパラメータを用いて、訓練フェーズで用いた検索対象データとは異なるデータに対して検索タスクを実行し、検索結果を評価する。

それぞれの実験フェーズごとに検索有効性は算出するが、従来の平滑化手法と提案手法の評価を行う際にはテストフェーズの評価値を用いる。

意見文検索タスクに与える検索対象データは、各実験フェーズに対して MPQA corpus を 2 等分した文集合を与えた。それぞれの実験フェーズで検索される文集合には、重複は存在しない。この

設定により、検索対象データに強く依存せずに検索有効性のよいパラメータを設定したとき、提案手法の適用によりどの程度精度の改善がみられるかを評価する。

5.2 評価指標

実験の評価で用いた評価指標は、Bpref [13] と MAP [14] である。Bpref は、文の適合判定が完全に付加されていない場合でも、古典的な指標と比較して安定な評価を行うことができる指標である。また、MAP (Mean Average Precision) は、検索結果の精度に着目した指標である。

本実験では、第 5.1 節でも述べたように話題の適合判定と意見の適合判定のそれぞれに対して仮定をおいている。そのため、必ずしも完全な適合判定が行われるとは限らない。そこで実験の評価では Bpref により着目することにする。

5.3 実験結果と考察

5.3.1 実験結果

まず 1 つ目の実験として、局所文脈を用いた文検索の精度をより高めるために、局所文脈の幅の最適値を求める実験を行った。話題に関する局所

Table 2: The results of main experiments with Bpref. For each seed word set, left values are results of proposed method, and right values in parentheses are baseline results.

Models	KAM		TUR		LAV	
lmtf	0.2197	(0.1385)	0.2197	(0.1385)	0.2197	(0.1385)
basef	0.2171	(0.1416)	0.2137	(0.1451)	0.2046	(0.1550)
rmtf	0.2911	(0.1938)	0.2911	(0.1938)	0.2911	(0.1938)
rmt-basef	0.2844	(0.1825)	0.2807	(0.1976)	0.2766	(0.1957)
rms-basef	0.2749	(0.1326)	0.2592	(0.1301)	0.2391	(0.1268)
slmf	0.2885	(0.2044)	0.2838	(0.2055)	0.2794	(0.2031)

Table 3: The results of main experiments with MAP. For each seed word set, left values are results of proposed method, and right values in parentheses are baseline results.

Models	KAM		TUR		LAV	
lmtf	0.1799	(0.1119)	0.1799	(0.1119)	0.1799	(0.1119)
basef	0.1792	(0.1142)	0.1746	(0.1188)	0.1852	(0.1369)
rmtf	0.2983	(0.1776)	0.2983	(0.1776)	0.2983	(0.1776)
rmt-basef	0.3280	(0.1704)	0.3240	(0.1882)	0.3195	(0.1862)
rms-basef	0.2792	(0.1158)	0.2623	(0.1148)	0.2441	(0.1096)
slmf	0.3172	(0.1698)	0.3162	(0.1668)	0.2843	(0.1714)

文脈と意見に関する局所文脈を一致させた場合の実験結果を Fig. 1, Fig. 2 に示す。また、Bpref と MAP のそれぞれの評価指標を用いて評価を行った一覧を Table. 1 に示す。

Fig. 1, Fig. 2 によると、局所文脈の幅を前後 1 文から前後 5 文に変更していくことで、検索有効性の急激な向上がみられた。

また、Table. 1 によると、Bpref で評価を行った場合においてシードワード KAM では局所文脈の幅を前後 30 文に設定した時が最も評価値が高く、シードワード TUR, LAV では前後 20 文に設定した時が最も評価値が高くなった。MAP で評価を行った場合においてはシードワード KAM, LAV では前後 20 文に設定した時が最も評価値が高く、シードワード TUR では前後 30 文に設定した時が最も評価値が高くなった。このような結果から、前後 20 文から 30 文を局所文脈として設定した場合に最も検索有効性が高くなると言える。また、局所文脈の幅を無限大に設定した場合においても総じて高い評価値を得ている。

局所文脈を用いて平滑化を行った場合 (Table. 1 の width が 0 の場合) と、局所文脈を用いずに平滑化を行った場合 (Table. 1 の width が 1 の場合) とでは、Bpref で評価した場合は 12.3 % から 20.2 % の改善率を得た。MAP で評価した場合は 12.2 % から 22.9 % の改善率を得た。

次に、最適な局所文脈の幅を設定した上で提案手法と従来手法 [3] の実験結果を比較した。その結果を Table. 2, Table. 3 に示す。クエリ尤度モ

デルを用いた検索タスク lmtf, basef では、Bpref で評価した場合、32.0 % から 58.8 % の改善率を得た。また、MAP で評価した場合は 35.3 % から 60.7 % の改善率を得た。

適合モデルを用いた検索タスク rmtf, rmt-basef, rms-basef, slmf では、Bpref で評価した場合、37.5 % から 110.7 % の改善率を得た。また、MAP で評価した場合は 65.86 % から 141.1 % の改善率を得た。

検索タスクの中では、特に rms-basef での改善が強くみられた。また、クエリ尤度モデルを用いた検索タスクよりも適合モデルを用いた検索タスクの方が、改善の度合いが大きく現れた。

5.3.2 考察

話題に関する局所文脈の幅と意見に関する局所文脈の幅を一致させた実験では、設定した検索タスク全てで改善が見られた。これは、適合文の付近には別の適合文が存在することを証明した形であるともいえる。また、局所文脈の幅を前後 1 文から前後 5 文まで広げていくと急激に評価値が向上したのは、ある適合文の前後 5 文の範囲に別の適合文が存在する可能性が高いことを示している。

また、クエリ尤度モデルよりも適合モデルでの改善の度合いが高いのは、適合モデルで用いられるクエリが従来手法よりも適合クラスに属する割合が増加したからと考えられる。適合の割合が高いクエリが適合モデルに与えられることで、より適合文を検索する性能が向上したといえる。

6 おわりに

本論文では、意見文検索における検索有効性を高めるための言語モデルの平滑化手法を提案した。文、局所文脈、文書集合の確率的言語モデルを平滑化した混合モデルを用いて、従来の意見文検索手法と同様の実験環境によって比較実験を行った。これにより、Bpref による評価結果ではクエリ尤度モデルの枠組みで 32.0 % から 58.8 %, 適合モデルの枠組みで 65.86 % から 141.1 % の検索有効性の改善率を得た。

今後の課題として、本実験での適合判定に仮定をおかない場合において、提案手法がどの程度の成果を挙げるかの検討が挙げられる。

参考文献

- [1] Hu, M. and Liu, B.: Mining and summarizing customer reviews, in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 168–177 (2004).
- [2] Kim, S. M. and Hovy, E.: Determining the sentiment of opinions, in *Proceedings of the Conference on Computational Linguistics* (2004).
- [3] Eguchi, K. and Lavrenko, V.: Sentiment Retrieval using Generative Models, in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pp. 245–254 (2006).
- [4] Ponte, J. M. and Croft, W. B.: A language modeling approach to information retrieval, in *Proceedings of the 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp. 275–281 (1998).
- [5] Song, F. and Croft, W. B.: A General Language Model for Information Retrieval, in *Proceedings of the 8th International Conference on Information and Knowledge Management* (1999).
- [6] Lavrenko, V. and Croft, W. B.: Relevance Based Language Models, in *Proceedings of the 24th Annual International ACM SIGIR Conference*, pp. 120–127 (2001).
- [7] MacKay, D. J. C. and Peto, L. C. B.: A hierarchical Dirichlet language model, *Natural Language Engineering*, Vol. 1, No. 3, pp. 1–19 (1995).
- [8] Wilson, T., Wiebe, J. and Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis, in *Proceedings of Human Language Technologies Conference* (2003).
- [9] Wiebe, J., Wilson, T. and Cardie, C.: Annotating expressions of opinions and emotions in language, *Language resources and Evaluation*, Vol. 39, No. 2–3, pp. 165–210 (2005).
- [10] Kamps, J. and Marx, M.: Words with attitude, in *Proceedings of the 1st International Conference on Global WordNet*, pp. 332–341 (2002).
- [11] P. D. Turney, M. L. L.: Measuring praise and criticism: Inference of semantic orientation from association, *ACM Transactions on Information Systems*, Vol. 21, No. 4, pp. 315–346 (2003).
- [12] Krovetz, R.: Viewing morphology as an inference process, in *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 191–202 (1993).
- [13] Buckley, C. and Voorhees, E. M.: Retrieval evaluation with incomplete information, in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 25–32 (2004).
- [14] Baeza-Yates, R. and Ribeiro-Neto, B.: *Modern Information Retrieval*, Addison Wesley (1999).