Vol.2009-DBS-148 No.16
Vol.2009-FI-95 No.16
2009/7/28

意見文検索のための言語モデルにおける局所文脈スムージング

岡本和剛[†1]　本田徹也[†1]　江口浩二[†1]

近年，Web の発展と普及に伴って情報の発信が容易になり，膨大な情報がアクセス可能になった．このような情報を検索する手法の中で重要なものの一つに意見文検索がある．ある着目する話題に対する肯定的もしくは否定的な意見を含む文を検索する機能の有用性は高い．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，局所文脈なる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用いられる従来の言語モデルのスムージング手法を拡張し，局所文脈に関する言語モデルを用いて拡張した新たな手法を提案する．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 of Language Models for Sentiment Sentence Retrieval

Takayoshi Okamoto,[†1] Tetsuya Honda[†1] and Koji Eguchi[†1]

Recently a number of documents are published on the web. One of the crucial techniques to access to such information 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 retrieving sentiment-bearing sentences. 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 神戸大学大学院工学研究科情報知能学専攻
Graduate School of Engineering, Kobe University

1. はじめに

近年，Web の発展と普及に伴って情報の発信が容易になり，膨大な情報がアクセス可能になった．このような文書または文を検索する技術の高度化が求められているところであるが，とりわけ，本研究では意見文検索に着目する．意見文検索とは，ある話題に対して肯定的な文もしくは否定的な文を検索要求に対する適合性の度合に従ってランキングする技術を指す．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，あるいは，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る．

意見情報検索と類似するタスクに意見極性分類がある．意見極性分類に関する一部の研究においては意見極性の話題依存性が着目されている[1)–4)]．一方，意見情報検索の研究は比較的新しく，意見表現や意見極性の話題依存性を考慮した研究事例は多くない．このような研究に文献 5) がある．(1) 検索要求の意見的側面と話題的側面を区別し，検索対象テキストデータの意見部分とそれ以外の部分に対するそれぞれの類似度を求めてそれらを結合した初期検索を行い，(2) さらにその結果を用いて検索要求のモデルを推定し検索に用いることで，意見の話題依存性を考慮している．さらに，意見の対象を表すような検索語を与えるだけでなく，肯定，否定などの意見極性を指定した，極性指定型意見情報検索なるタスクが最初に定義されている．しかしながら，意見文検索の検索有効性という観点からは依然として改善の余地がある．

そこで本稿では，文書中のある局所的範囲では意見極性が一貫していることが多いという特性に着目し，局所文脈なる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．本稿では，確率的言語モデルに基づく文書検索に用いられるディリクレ・スムージング[6)] を拡張し，文レベルの検索を想定した，局所文脈に基づくスムージング手法を提案する．本稿では，提案するスムージング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検索有効性が大きく改善されることを実験によって示す．

2. 意見文検索

本節では，まず文検索に適用可能な既存の検索モデルについて概説する．次に，意見文検索のために提案された検索モデルについて述べる．これらは後節において，提案手法によるスムージングと組み合わせる．

2.1 文検索

本稿では，文検索を，対象となる文書コレクションからユーザのもつ情報要求に適合する

Vol.2009-DBS-148 No.16
Vol.2009-FI-95 No.16
2009/7/28

文を検索する処理を指すものとする．入力されたクエリをもとに推定された情報要求と文との適合度によって文をランキングし，ユーザに提示する．以下に，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一つに，多項分布などで表現された文書モデルからクエリ語が生成される尤度を適合度とみなす検索モデルがある．このような検索モデルを**クエリ尤度モデル**[7)–9)] と呼び，文書を単位とした検索手法として 1998 年以降に提案されている．この検索モデルにおいて文書モデルを文モデルに置き換えることで，文検索を実現することができる．このとき，文 S のモデルからクエリ $Q = (q_1, \ldots, q_n)$ が生成される尤度 $P(Q|S)$ は式 (1) のように定義され，これに基づいて文をランキングする．

$$P(Q|S) = \prod_{i=1}^{n} P(q_i|S) = \prod_{w \in Q} P(w|S)^{c(w,Q)} \tag{1}$$

ここで $c(w, Q)$ は，クエリ Q の中で語 w が出現する頻度である．文モデル $P(w|S)$ は次式のような最尤推定すなわち文中の語の相対頻度によって推定する．

$$P(w|S) = P_{ml}(w|S) = \frac{c(w, S)}{|S|} \tag{2}$$

式 (1)，式 (2) より，文中に出現しない，すなわち文モデルにおいて非零の確率値が割り振られていないクエリ語 w が一つでも存在すると，他にどのようなクエリ語が存在していようとも尤度 $P(Q|S)$ が 0 になることに注目されたい．このように尤度が 0 となり他の文との比較が困難になる状況を零確率問題という．零確率問題を回避する手法の一つとして，言語モデルのスムージングがよく用いられる．これについては 3.1 節で説明する．なお，スムージングの適用によって零確率問題が解消された場合，式 (1) による文のランキングは，次式で得られるクエリ Q の言語モデルと文 S の言語モデルとの負のクロスエントロピー $-H(Q||S)$ による文ランキングと等価である．

$$-H(Q||S) = \sum_{w \in Q} P(w|Q) \log P(w|S) \tag{3}$$

2.1.2 適合モデル

仮に検索対象の文書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**[10)] は適合クラスに属する文書群を用いて推定された，情報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．前節で述べたクエリ尤度モデルの場合と同様に，ここでも，文書を文に置き換えることで文レベルの検索を実現できる．

クエリを $Q = (q_1, \ldots, q_n)$，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．

$$P(w|R) = P(w|q_1, \ldots, q_n) = \frac{P(w, q_1, \ldots, q_n)}{P(q_1, \ldots, q_n)} \tag{4}$$

ここで，クエリ $(q_1, \ldots, q_n)$ 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．

文書コレクション C からある文 S を選択する確率を $P(S)$ とするとき，文 S から $n+1$ 回語を選択する場合を考える．クエリ $(q_1, \ldots, q_n)$ と共に語 w を選択する確率を式 (5) に示す．

$$P(w, q_1, \ldots, q_n) = \sum_{S \in C} P(S) P(w, q_1, \ldots, q_n|S) \tag{5}$$

クエリ $(q_1, \ldots, q_n)$ と語 w はそれぞれ独立に選択しているため，式 (5) の $P(w, q_1, \ldots, q_n|S)$ は次のようにも表わすことができる．

$$P(w, q_1, \ldots, q_n|S) = P(w|S) P(q_1, \ldots, q_n|S) \tag{6}$$

$$= P(w|S) \prod_{i=1}^{n} P(q_i|S) \tag{7}$$

式 (6) を式 (5) に代入すると，次のように表わせる．

$$P(w, q_1, \ldots, q_n) = \sum_{S \in C} P(S) P(w|S) \prod_{i=1}^{n} P(q_i|S) \tag{8}$$

これにより，式 (8) を式 (4) に代入することで，次の式を導ける．

$$P(w|R) = \frac{\sum_{S \in C} P(S) P(w|S) \prod_{i=1}^{n} P(q_i|S)}{P(q_1, \ldots, q_n)} \tag{9}$$

上式における $\prod_{i=1}^{n} P(q_i|S)$ は，式 (1) に示したクエリ尤度モデルによって求めることができる．

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すこともできる．

$$P(w|R) = \sum_{S \in C} P(w|S) P(S|q_1, \ldots, q_n) \tag{10}$$

現実的には適合モデル $P(\cdot|R)$ を次のような近似によって求める．すなわち，$P(S|q_1, \ldots, q_n)$

Vol.2009-DBS-148 No.16
Vol.2009-FI-95 No.16
2009/7/28

の降順に順序付けされた文の上位 N 個を利用して，式 (10) に従って混合分布を構築し，この混合分布において $P(w|S)P(S|q_1,\ldots,q_n)$ の値が大きい語 w のうち M 語を用いることで，$P(\cdot|R)$ を近似する．

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類似度として負の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 V とした時，適合クラス R の言語モデルと文 S の言語モデルとの負のクロスエントロピー $-H(R||S)$ は次のように定義される．

$$-H(R||S) = \sum_{w \in V} P(w|R) \log P(w|S) \tag{11}$$

式 (11) から，$-H(R||S)$ の降順に文を順序付ける．

以上に述べた適合モデルは疑似適合フィードバック[11] の言語モデルによる実現と見なすことができ，同義語と多義語の問題を軽減する効果がある．数々の実証実験において，適合モデルはクエリ尤度モデルと比較して検索有効性が高いことが報告されている[10]．

2.2 意見文検索

本節では，2.1.1 と 2.1.2 で紹介した文検索のための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意見クエリ尤度モデルと話題・意見適合モデルについて文献 5) に基づいて説明する．

2.2.1 意見クエリ尤度モデル

ある言語モデル $\mathbf{p_1}$，$\mathbf{p_2}$ と意見極性モデル $\mathbf{p_x}$ に対して確率質量関数 $\pi(\mathbf{p_1}, \mathbf{p_2}, \mathbf{p_x})$ を定義する．ある文 $S_i = (w_{i1}, \ldots, w_{im})$ 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 2 値変数 $b_{ij} \in \{S, T\}$ 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，文 $S_i = (w_{i1}, \ldots, w_{im})$ とその意見極性 x_i を観測する確率は次のように表せる．

$$\sum_{\mathbf{p}_t, \mathbf{p}_s, \mathbf{p}_x} \pi_i(\mathbf{p}_t, \mathbf{p}_s, \mathbf{p}_x) \mathbf{p}_x(x_i) \prod_{j=1}^{m} \begin{cases} \mathbf{p}_t(w_{ij}) & (\text{if } b_{ij} = T) \\ \mathbf{p}_s(w_{ij}) & (\text{otherwise}) \end{cases} \tag{12}$$

ただし，$\mathbf{p}_t$ は話題に関する言語モデル，$\mathbf{p}_s$ は意見に関する言語モデルである．ここで，説明の便宜上，指示関数 $\delta(y)$ を定義する．指示関数 $\delta(y)$ は，述語 y が真であれば 1，そうでなければ 0 を返す．このとき，関数 $\pi_i(\mathbf{p}_t, \mathbf{p}_s, \mathbf{p}_x)$ を，言語モデル $(\mathbf{p}_t, \mathbf{p}_s, \mathbf{p}_x)$ が文 S_i のモデル $(\mathbf{p}_{ti}, \mathbf{p}_{si}, \mathbf{p}_{xi})$ に対応するときに $\frac{1}{n}$ を返し，そうでないとき 0 を返す確率質量関数として定義する．

$$\pi_i(\mathbf{p}_t, \mathbf{p}_s, \mathbf{p}_x) = \frac{1}{n} \sum_{i'=1}^{n} \delta(\mathbf{p}_t = \mathbf{p}_{ti'}) \delta(\mathbf{p}_s = \mathbf{p}_{si'}) \delta(\mathbf{p}_x = \mathbf{p}_{xi'}) \tag{13}$$

式 (12) による検索モデルを**意見クエリ尤度モデル**と呼ぶ．

なお，すでに我々は文献 12) で，意見に関するシードワードを用いるアプローチにおける局所文脈の影響について検討した．本稿で後述する実験では，意見シードワードの代わりに，意見極性が付与された訓練データを用いる状況[5] を仮定する．この条件下では，上述の意見クエリ尤度モデルは，検索対象のテキストデータに意見極性が付与された状況を仮定する．そのため，このモデル単独では実用的とは言えないが，次節で述べる話題・意見適合モデルの推定に利用できる．

2.2.2 話題・意見適合モデル

意見クエリ尤度モデルに基づいて適合モデルを拡張することで意見文検索に応用できる．本節においてその形式的な定義について述べる．文書コレクション C とクエリ $(\mathbf{q}_t, \mathbf{q}_s, \mathbf{q}_x)$ が与えられたとする．ここで，$\mathbf{q}_t$ を話題を表すクエリ語，$\mathbf{q}_x$ を着目する意見極性，$\mathbf{q}_s$ を着目する意見極性に対応したシードワードであるとする．意見に依存した話題に関する適合モデル $\mathbf{r}_t$ と話題に依存した意見に関する適合モデル $\mathbf{r}_s$ は以下のように推定する．クエリ $(\mathbf{q}_t, \mathbf{q}_s, \mathbf{q}_x)$ が式 (12) で表された確率分布からランダムにサンプリングされたと仮定し，さらにもう 1 つの話題または意見に関する語 w がサンプリングされる尤度は次のようになる．

$$\mathbf{r}_t(w) = \frac{P(\mathbf{q}_s, \mathbf{q}_t \circ w, \mathbf{q}_x)}{P(\mathbf{q}_s, \mathbf{q}_t, \mathbf{q}_x)} \quad , \quad \mathbf{r}_s(w) = \frac{P(\mathbf{q}_s \circ w, \mathbf{q}_t, \mathbf{q}_x)}{P(\mathbf{q}_s, \mathbf{q}_t, \mathbf{q}_x)}. \tag{14}$$

ここで，$\mathbf{q} \circ w$ は語の列 $\mathbf{q}$ に語 w を追加する操作を示す．

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ランキングする．従来の適合モデルではクロスエントロピーを用いてランキングを行っていたが，ここではそれを次式のように変形してランキングに用いる．

$$\nu \sum_{w \in V} \mathbf{r}_t(w) \log \mathbf{p}_{ti}(w) + (1 - \nu) \sum_{w \in V} \mathbf{r}_s(w) \log \mathbf{p}_{si}(w) \tag{15}$$

ここで，V を文書コレクションに含まれる語彙集合とし，$\mathbf{p}_{ti}$ と $\mathbf{p}_{si}$ はそれぞれ文 S_i の話題部分と意見部分の言語モデルを示し，最尤推定によって推定する．ν は話題と意見とのバランスを調整するパラメータを示し，開発データを用いて経験的に定める．この検索モデルを**話題・意見適合モデル**と呼ぶ．

さて，2.2.1 で言及した通り，本稿で後述する実験では，求める意見極性 $\mathbf{q}_x$ と話題語 $\mathbf{q}_t$ のみが与えられる状況を仮定し，意見極性付き訓練データを用いて話題・意見適合モデルを推

Vol.2009-DBS-148 No.16
Vol.2009-FI-95 No.16
2009/7/28

定する．このとき，式 (14) において $\mathbf{q}_s = \emptyset$ すなわち空集合と考えればよい．つまり，訓練データを用いて話題・意見適合モデルを推定し，開発データを用いてパラメータを推定した後，意見極性の付与されていないテストデータに対して検索を実行する．

3. 局所文脈を用いたスムージング

3.1 ディリクレ・スムージング

文書検索のための言語モデルのスムージング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) 文書モデルにおいてあらゆる語彙に非零の微小な確率値を割り振ることにより零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．

スムージングの研究は多くなされており，様々な手法が提案されている．なかでも，文書検索の目的で特に有効であると報告されているのがディリクレ・スムージング[6] である．以下ではこれについて説明する．

ディリクレ・スムージングは，線形補間法（Jelinek-Mercer スムージング）[13] に対して文書長すなわち文書の述べ語数を加味した手法である．線形補間法は，文書モデルを文書コレクションのモデルを用いて補間するものであるが，そのとき個々の文書の長さが大きく異なる場合でも同じ比率で補間する．そのため，文書長が小さい文書に対しては補間が不足し，逆に文書長が大きい文書に対しては補間しすぎる傾向がある．それに対して，ディリクレ・スムージングでは文書モデルとコレクションモデルの混合比を表わすスムージングパラメータに文書長を考慮した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もたらす影響を適度に調整し，文書長にばらつきのある文書コレクションに対しても有効性が期待される．

以上では文書モデルを想定したスムージングについて説明したが，文を対象とした場合に，最も単純な方法は従来のディリクレ・スムージングにおいて文書を文に置き換えることである．この場合の形式的な定義を与える．式 (2) で示した最尤推定によって得られた文モデル $P_{ml}(w|S)$ は，同じく最尤推定によって得られたコレクションモデル $P_{ml}(w|C)$ を用いてディリクレ・スムージングを適用すると式 (16) のようになる．

$$P(w|S) = \lambda P_{ml}(w|S) + (1-\lambda)P_{ml}(w|C) \tag{16}$$

$$\lambda = \frac{|S|}{|S| + \alpha} \tag{17}$$

ここで，$|S|$ は文 S を構成する語の述べ数である．また，α は 0 以上の値をとるスムージングパラメータで，その値は実験により経験的に定める．

3.2 意見文検索における局所文脈

一つの文書においては複数の話題または複数の意見が取り上げられることが少なくなく，とくに Web 上に公開されたブログなどの文書では，その傾向は顕著である．そのような文書から，指定した対象に関する意見を含む文を的確に抽出するためには，文脈を考慮して文の言語モデルを推定することが重要となる．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構成することが可能となる．

そこで本稿では，以下の**局所文脈**なる概念を導入する．局所文脈を，着目する文とその前後に存在する複数の文からなる集合として定義する．以下に局所文脈の言語モデルについて形式的な定義を与える．文書モデルを $D = \{S_1, S_2, \ldots, S_n\}$，着目する文モデルを S_k とする時，局所文脈 $L_k = L(k, \delta, D)$ の言語モデルを次のように定義する．

$$P(w|L_k) = \frac{\sum_{i=\max(1,k-\delta)}^{\min(n,k+\delta)} |S_i| P_{ml}(w|S_i)}{\sum_{i'=\max(1,k-\delta)}^{\min(n,k+\delta)} |S_{i'}|} \tag{18}$$

ここで，δ は文を単位とした局所文脈の幅を示すパラメータであり，非負の整数値をとる．局所文脈の幅 δ を十分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ため，上式の局所文脈モデルの定義は文書モデルを含む一般的なものである．

3.3 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，3.1 節で紹介したディリクレ・スムージングの単純な適用ではコレクションモデルのみを用いているため，そのままでは文検索に十分な効果が期待できないと考えられる．そこで，ディリクレ・スムージングを拡張し，文書コレクションと文の中間的な粒度の言語モデルをも用いたスムージング手法を提案する．

スムージングの適用対象である文を S_k とし，式 (2) で示した最尤推定によって得られた文モデルと文書コレクションモデルをそれぞれ $P_{ml}(w|S_k)$，$P_{ml}(w|C)$，そして，式 (18) で得られた局所文脈モデルを $P(w|L_k)$ とするとき，文モデル $P(w|S_k)$ の局所文脈スムージングを次のように定義する．

$$P(w|S_k) = \lambda P_{ml}(w|S_k) + (1-\lambda)P^*(w|L_k) \tag{19}$$

$$P^*(w|L_k) = \mu P(w|L_k) + (1-\mu)P_{ml}(w|C) \tag{20}$$

ここで，λ と μ は次のように定義する．

Vol.2009-DBS-148 No.16
Vol.2009-FI-95 No.16
2009/7/28

$$\lambda = \frac{|S|}{|S| + \alpha} \quad , \quad \mu = \frac{|L|}{|L| + \beta} \tag{21}$$

$|S|$ は文 S に含まれる延べ語数であり，$|L|$ は局所文脈 L 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，それぞれ 0 以上の値をとる．このスムージングパラメータは実験により経験的に定める．ただし，局所文脈幅 δ が 0 の場合は式 (19) の代わりに式 (16) を用いる．

4. 実験と評価

4.1 評価実験

提案手法の局所文脈スムージングの有効性を評価するために，意見文検索を想定したいくつかの実験を行う．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．

- 局所文脈スムージングにおいて検索有効性が最も向上する局所文脈幅はどの程度か．
- 局所文脈スムージングを用いることによって従来のスムージング手法と比較して検索有効性がどの程度向上したか．

評価実験では，文献 5) と同様に，MPQA Opinion Corpus version 1.2[4),14)]（以下，MPQA データ）を用いた．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．MPQA データでは，10 種類の話題について文書単位でラベルが付与されている．また，MPQA データには意見を示すフレーズに対してアノテーションが付与されている．そこで，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．

すでに上で述べた通り，元々の MPQA データには文書レベルで話題についてのラベルが付与されており，文レベルでの話題適合性が特定できない．そのため，我々は 10 種類の話題ラベルが付与された文書を構成する個々の文について，その話題に関する適合性を判定した．この話題適合性判定の結果と，前述の意見極性の仮定を組み合わせることで，文レベルでの話題と意見に関する適合性（以下，話題・意見適合性）を特定することが可能となった．

本稿では，10 種類の話題のそれぞれについて，肯定意見を求める場合と否定意見を求める場合を想定して，合計 20 通りの話題・意見適合性を仮定した極性指定型意見文検索の実験を行う．

さらに，意見文検索タスクに対して，以下の 3 つの実験フェーズを設定し評価する．

訓練フェーズ: 話題を表すクエリ語と意見極性付き訓練データを用いて話題・意見適合モデルを推定する．

開発フェーズ: 訓練フェーズで推定した話題・意見適合モデルを用いて，意見極性なし開発データに対して検索を実行し，検索有効性が最良となるように各種パラメータを決定する．

テストフェーズ: 決定されたパラメータのもとで意見極性付きの訓練データと開発データを用いて話題・意見適合モデルを推定し，それを用いて意見極性なしテストデータに対して検索を実行し，検索結果を評価する．

このため，元データをランダムに三等分し，それぞれを訓練データ，開発データ，テストデータとした．

なお，実験で MPQA データを用いるにあたり，文中に含まれる単語，またクエリ語に対して Krovetz stemmer[15)] を適用し，さらに 418 個のストップワード[16)] を除去した．

4.2 評価指標

実験の評価で用いた評価指標は，平均精度（Mean Average Precision，以下 MAP）[17)] と Bpref[17)] である．MAP は情報検索分野において広く受け入れられた標準的な評価指標であるため，実験の評価は主に MAP に基づくこととする．また，Bpref は適合判定が完全な形でなくとも安定した評価を行うことができると報告されている[17)]．この Bpref は，参考のため，MAP による評価結果に付記する．

4.3 実験結果

まず 1 つ目の実験として，局所文脈を用いた意見文検索の有効性をより高めるために，局所文脈幅の最適値を求めるための実験を行った．MAP と Bpref に基づく評価結果をそれぞれ図 1，2 に示す．また，これらに対応する評価値を表 1 に示す．ただし，以上の評価結果は開発フェーズでの評価結果である[⋆1]．なお，局所文脈幅に応じて式 (19) における μ などのパラメータの最適値が変わると考えられるため，以上の評価結果では，局所文脈幅ごとに他のパラメータを最適化したときの実験結果を用いた．

図 1，図 2 および表 1 によると，局所文脈幅 δ を前後 3 文から前後 5 文に増やした付近で，検索有効性の急激な向上がみられた．また，局所文脈幅として前後 7 文に設定した場合

表 1 局所文脈幅ごとの評価結果

局所文脈幅	0	1	3	5	7	10	15	20	25
MAP	0.0793	0.0843	0.0849	0.1028	0.1090	0.0900	0.0923	0.0888	0.0885
Bpref	0.0798	0.0849	0.0862	0.1037	0.1082	0.0967	0.0892	0.0838	0.0883

⋆1 これはテストデータを用いて局所文脈幅などのパラメータを調整することを避けるための措置である．

Vol.2009-DBS-148 No.16
Vol.2009-FI-95 No.16
2009/7/28

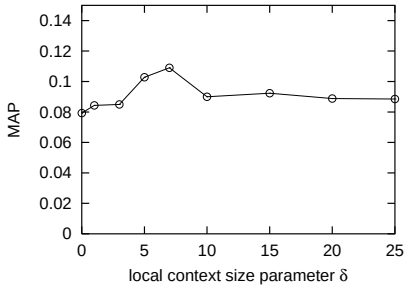

図 1 局所文脈幅と MAP の関係

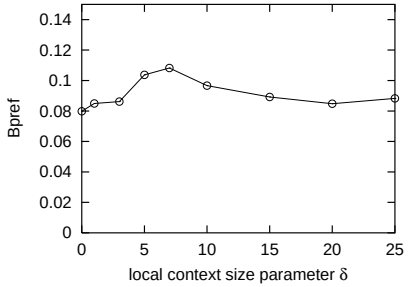

図 2 局所文脈幅と Bpref の関係

に評価値が最大であった．以上のことから，適切な局所文脈幅は前後 5 文から 7 文程度であると言える．

式 (19) によって局所文脈スムージングを行った場合（表 1 の局所文脈幅が 1 以上の場合）と，式 (16) によって局所文脈を用いずにスムージングを行った場合（表 1 の文脈の幅が 0 の場合）とを比較すると，局所文脈を用いることで MAP で評価した場合は 6.3% から最大 37.5% の，Bpref で評価した場合は 6.3% から最大 35.6% の改善率を得た．ここでいう改善率は (提案手法の評価値)/(比較対象の評価値)-1 で計算した．

次に，最適な局所文脈幅のもとで，提案する局所文脈スムージング（式 (19)）と従来のスムージング手法（式 (16)）による実験結果を比較した．その結果を表 2 に示す．

これは，テストフェーズすなわち局所文脈幅をはじめとして各種パラメータを開発データを用いて決定した上でテストデータに対して実施した実験結果である．提案する局所文脈

表 2 スムージング手法の評価

手法	従来手法（局所文脈幅=0）	提案手法（局所文脈幅=7）
MAP	0.0334	0.0927
Bpref	0.0358	0.0901

スムージングを用いることで MAP で評価した場合は 177.5% の，Bpref で評価した場合は 151.7% の改善率を得た．これは Wilcoxon 符号付順位検定（両側）を用いた検定において有意水準 0.05 の条件で有意な改善であることを確認した．

4.4 考　　察

以上の実験において，提案する局所文脈スムージング手法を用いることで従来のスムージング手法を用いる場合よりも検索有効性が改善されることを示すことができた．一般に，文書の先頭付近と末尾付近では話題が同じとは限らない．また，指示語がもたらず問題が起こり得る．例として，

あそこにいる女性はひろこです．
彼女はとてもきれいです．
実は，彼女は僕のいとこです．

という 3 つの文で構成される文書またはその一部分があったとする．第 2 文で出現する「彼女」は，第 1 文で出てくる「ひろこ」を指している．また,「とてもきれい」が修飾しているのは「彼女」である．従来の手法だと 1 つの文しか見ないため，第 2 文の「きれい」という意見表現の対象が何であるかが特定できない．このことから，意見文検索において，求める結果が得られない可能性が高い．局所文脈を導入することによって，以上のような問題が緩和され，有効な意見文検索が実現すると考えられる．

また，図 1，図 2 および表 1 では局所文脈幅を 0 から 25 文を範囲として実験を行っている．前後 25 文をとると局所文脈を構成する文の数は 51 になり，本実験で用いた MPQA データでは 1 つの文書全体を指す場合がほとんどである．この場合，局所文脈を前後 7 文や 5 文とる場合に比べて検索有効性は悪くなっている．

1 つの文書において，複数の話題や複数の意見が含まれていることが少なくない．文書全体を局所文脈として用いると，話題においても意見においても正確な検索ができないため，最適な局所文脈幅（本実験では前後 7 文）の場合よりは検索有効性が劣る．これが理由であると考えられる．ただし，それでも従来手法よりは検索有効性が改善していることは意見文検索において文脈を考慮することがいかに重要であるかを示していると言える．

Vol.2009-DBS-148 No.16
Vol.2009-FI-95 No.16
2009/7/28

5. おわりに

本稿では，意見文検索における検索有効性を改善するための言語モデルのスムージング手法を提案した．文，局所文脈，文書コレクションのユニグラム言語モデルからなる混合モデルを構成することにより，文モデルの効果的なスムージングを実現した．評価実験では，話題・意見適合モデル[5] なる意見文検索モデルに対して，提案するスムージング手法を適用し，従来型のスムージング手法を適用した場合と比較した．

これにより，平均精度による評価結果において，提案手法の局所文脈を考慮したスムージング手法は，局所文脈を導入しない従来のスムージング手法と比べて 177.5% の有意な改善率を得た．これにより，意見文検索のためのスムージングに局所文脈を導入することの効果を示すことができた．大規模かつ多様なデータコレクションを用いて，局所文脈の適切な粒度などに関する詳細な調査を行うことが今後の課題である．

参考文献

1) Nasukawa, T. and Yi, J.: Sentiment Analysis: Capturing Favorability using Natural Language Processing, *Proceedings of the 2nd International Conference on Knowledge Capture (K-CAP-03)*, Sanibel Island, Florida, USA, pp.70–77 (2003).
2) Yi, J., Nasukawa, T., Bunescu, R. and Niblack, W.: Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques, *Proceedings of the 3rd IEEE International Conference on Data Mining*, Melbourne, Florida, USA, pp. 427–434 (2003).
3) Engström, C.: Topic Dependence in Sentiment Classification, Master's thesis, University of Cambridge (2004).
4) Wilson, T., Wiebe, J. and Hoffmann, P.: Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis, *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, Vancouver, British Columbia, Canada (2005).
5) Eguchi, K. and Lavrenko, V.: Sentiment Retrieval using Generative Models, *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, pp.345–354 (2006).
6) Zhai, C. and Lafferty, J.: A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval, *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New Orleans, Louisiana, USA, pp.334–342 (2001).
7) Ponte, J.M. and Croft, W.B.: A Language Modeling Approach to Information Retrieval, *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, pp.275–281 (1998).
8) Hiemstra, D.: A Linguistically Motivated Probabilistic Model of Information Retrieval, *Research and Advanced Technology for Digital Libraries*, Lecture Notes in Computer Science, Vol.1513, Springer-Verlag, pp.569–584 (1998).
9) Song, F. and Croft, W.B.: A General Language Model for Information Retrieval, *Proceedings of the 8th ACM International Conference on Information and Knowledge Management*, Kansas City, Missouri, USA, pp.316–321 (1999).
10) Lavrenko, V. and Croft, W.B.: Relevance Based Language Models, *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New Orleans, Louisiana, USA, pp.120–127 (2001).
11) Buckley, C., Salton, G. and Allan, J.: Automatic Retrieval With Locality Information Using SMART, *Proceedings of the First Text REtrieval Conference (TREC-1)* (Harman, D.K., ed.), NIST Special Publication 500-207, pp.59–72 (1993).
12) 本田徹也，江口浩二：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，情報処理学会研究報告， No.2008-NL-184, pp.83–90 (2008).
13) Jelinek, F. and Mercer, R.L.: *Interpolated Estimation of Markov Source Parameters from Sparse Data*, pp.381–397, North-Holland Publishers (1980).
14) 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).
15) Krovetz, R.: Viewing Morphology as an Inference Process, *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Pittsburgh, Pennsylvania, USA, pp.191–202 (1993).
16) Callan, J.P., Croft, W.B. and Harding, S.M.: The INQUERY Retrieval System, *Proceedings of the 3rd International Conference on Database and Expert Systems Applications*, Valencia, Spain, pp.78–83 (1992).
17) Baeza-Yates, R. and Ribeiro-Neto, B.(eds.): *Modern Information Retrieval*, chapter3: Retrieval Evaluation, pp.73–97, Addison-Wesley (1999).