

主観的トリガー言語モデルによる意見情報検索

関 広[†] 上 原 邦 昭^{††}

本研究では、主観的意見は主たる二つの構成要素を持つことに着目し、統計的言語モデルの一つであるトリガー言語モデルを用いて意見情報検索を行うための枠組みを提案する。二つの構成要素の一つは、主観的意見を持つ主体または批評の対象であり、もう一つは意見を表す表現である。これをそれぞれトリガー、非トリガーと見なし、意見を記述した大量の文章からこれらの組（トリガー対）を発見する。本研究では、これらのトリガー対を基に主観的言語モデルを構築し、情報検索システムの再順位付けに利用する。評価実験によって、本提案モデルによって意見情報検索の性能を大きく向上させることができることを示す。

Subjective Trigger Models for Opinion Retrieval

KAZUHIRO SEKI[†] and KUNIAKI UEHARA^{††}

This paper presents a novel application of trigger language models to opinionated blog post retrieval. We employ the models to capture the characteristics of written personal opinions based on an assumption that there are two constituents to form a subjective opinion; One is the subject of the opinion or the object that the opinion is about, and the other is a subjective expression. The former is regarded as a triggering word and the latter as a triggered word. We automatically identify those subjective trigger patterns to build a language model from a large number of product customer reviews. Our experiments show that, when used for reranking initial search results, our proposed model significantly improves opinionated document retrieval by over 20% in MAP.

1. はじめに

現在ウェブ上には、ホームページや掲示板、ブログ等、ユーザが発信する様々な形態のコンテンツが存在する。中でも、ブログはその簡易性から、政治や趣味等に関する私的な意見・経験を表現するための方法として、多くのユーザに利用されている。本研究では、ブログを対象として、所与のエントリに関する主観的な意見を含んだ文書を検索する「意見情報検索」の手法を提案する。

意見情報検索の既存手法は、大まかに、辞書に基づく手法¹⁾と分類に基づく手法²⁾に分けることが出来る。前者は、人手で、あるいは自動で作成した主観的表現（例えは「like」や「fantastic」）のリストを用い、それらの語が現れる文書は主観的な文書であると見なす。一方、後者は、正例（意見を含む文書）と負例（意見を含まない文書）を基に機械学習の手法によって分類器を作成し、意見文書を同定する。

本稿では、どちらのカテゴリにも属さない新しい意見情報検索の手法について議論する。本手法は、代名詞あるいは主観的表現と批評の対象との文書中での距離を考慮することで、意見情報検索性能が向上する³⁾という経験的発見に部分的に着想を得たものであり、統計的言語モデルの一つであるトリガーモデル^{4),5)}を利用することで、主観的意見の記述に現れる特徴的なパターンを捉える。

2. 意見情報検索のための主観的言語モデル

2.1 トリガーモデル

一般に広く用いられている n グラム言語モデルは、 n (語) で示される局所的文脈しか考慮することが出来ない。より広い文脈を考慮するため、Lau ら⁴⁾はトリガーモデルを提案した。「トリガー」とは他の語の生起を誘発する語のことであり、トリガーと依存関係を持って生起する語を「非トリガー」と呼ぶ。また、これらの語の対を「トリガー対」と呼ぶ。このようなトリガー対の依存関係を表したトリガーモデル $P_T(w|h)$ は、通常、ベースラインの n グラム言語モデル $P_B(w|h)$ と線形結合して用いられる。

$$P_E(w|h) = (1 - \lambda) \cdot P_B(w|h) + \lambda \cdot P_T(w|h) \quad (1)$$

† 神戸大学自然科学系先端融合研究環

Organization of Advanced Science and Technology,
Kobe University

†† 神戸大学大学院工学研究科

Graduate School of Engineering, Kobe University

ここで, w と h は, それぞれ単語と履歴 (w に先行する語) を示し, λ は両言語モデルの影響を制御するパラメタである. ($P_T(w|h)$ の定義は 2.2 節で述べる.)

トリガーモデルを構築するためには, まず最初に(意味のある) トリガー対を同定する必要がある. Tillmann と Ney⁵⁾ は, n グラム言語モデルにおける語 w の生起確率 $P(w|h)$ がある閾値 t より小さいときだけ w を非トリガー候補とする基準(「低レベルトリガー」と呼ばれる)を提案した. この条件式を満足する全ての語 b に関して, あるトリガー対「 $a \rightarrow b$ 」のみによって拡張したトリガーモデル $P_{E:a \rightarrow b}(\cdot)$ と通常の n グラム言語モデル $P(\cdot)$ との対数尤度差を評価し, 依存関係にあるトリガー対を同定する. なお, a はコーパス中の全ての語である. 対数尤度差が大きい対から順に任意の数のトリガー対を採用することで, 最終的なトリガーモデル $P_T(\cdot)$ を得る.

本研究では, 上述のトリガーモデルを基に, 二種類の仮定を導入することで主観的意見に特徴的な語の依存関係を捉える. 一つ目の仮定は, 主観的意見は二つの基本的な構成要素, すなわち意見の主体または意見の対象と主観的表現を持つことである. 本研究では, 前者をトリガー, 後者を非トリガーとして扱う. 二つ目の仮定は, トリガーは多くの場合, 代名詞として表出することである. これらの仮定は, 代名詞または主観的表現と意見の対象との共起が, 意見らしさを示す手がかりとなるという経験的事実に拠っている³⁾.

従来の研究で用いられてきたアドホックな経験的事実と比較し, 本提案モデルは, 意見らしさを示す語間の依存関係をより理論的な枠組みによってモデルを取り込むことが出来る. また, 少数の代名詞のみを潜在的なトリガーとして扱うことで, より効率的かつ効果的に主観的意見に特化した言語モデルを構築することが出来る. 以下では, このような主観的トリガーモデルによって拡張された提案言語モデル $P_E(\cdot)$ を「主観的言語モデル」と呼ぶ.

2.2 主観的言語モデルの構築

前節の手順と仮定によって, 主観的言語モデルを以下のように構築する. まず, 主観的意見に特徴的なトリガー対を同定するため, amazon.com から 5,000 の顧客レビューを収集し, 実験に利用した. これらのレビューは, アマゾンで販売される様々な製品を対象に書かれたものである. また, 潜在的なトリガーとして, 本研究では次の 14 の代名詞を用いた.

I, my, you, it, its, he, his, she, her, we,
our, they, their, this

これらの設定のもと, 低レベルトリガーを用いて 10,000 のトリガー対を同定した結果の一部を表 1 に示す. なお, 履歴 h は同一文内の先行語とした.

表 1 が示すように, 非トリガーは機能語が多数を占め, 意見表現を特徴付けるようなトリガー対はほとんど同定されなかった. この問題は, 機能語は多くの文

表 1 対数尤度差が最大の低レベルトリガー対.

トリガー (a)	非トリガー (b)	$\Delta_{a \rightarrow b}$
this	→ the	7.079
it	→ the	7.079
i	→ the	7.079
i	→ to	6.526
this	→ to	6.525
my	→ and	6.502
	...	

脈に現れること, および低レベルトリガーは履歴 h の頻度を考慮していないことに起因する. そのため, 本研究では低レベルトリガーの条件式を $\tau \cdot P(w_i|h_i) < t$ のように修正する. ここで, τ は w_i と h_i の頻度の比と定義する. この修正によって, 低頻度の履歴 h_i を持つ高頻度の語 w_i にはペナルティが課され, 機能語等が非トリガーとして同定されにくくなる.

表 2 に, 修正した基準によって同定されたトリガー対を示す. 今回は, 意見表現を特徴づけるトリガー対が同定できていることが確認できる. また, 「I → very」や「it → greatest」, 「I → fantastic」のように, 遠距離の語の依存関係も抽出できていた.

表 2 修正後の基準を用いたときの対数尤度差が最大の低レベルトリガー対.

トリガー (a)	非トリガー (b)	$\Delta_{a \rightarrow b}$
i	→ wish	5.113
i	→ felt	5.073
i	→ loved	4.862
i	→ hope	4.739
i	→ couldn	4.680
i	→ got	4.611
i	→ cannot	4.593
	...	

続いて, 同定されたそれぞれのトリガー対 $(a \rightarrow b)$ について, それらの関連度 $\alpha(b|a)$ を共起関係に基づく最尤推定によって算出し, トリガーモデルを次のように構築する.

$$P_T(w|h) = \frac{1}{|h|} \sum_{w_j \in h} \alpha(w|w_j) \quad (2)$$

ここで, w_j は w の履歴に含まれる語である.

なお, ベースラインの言語モデル $P_B(w|h)$ としては, 平滑化したバックオフ 3 グラムモデルを用いた. また, 式 (1) における補間係数 λ は経験的に 0.9 として, 主観的言語モデルを構築した.

2.3 モデル適応

主観的言語モデルは製品のレビューから構築されたため, 製品以外に関する意見を同定するためにはあまり効果的ではないと考えられる. この問題に対処するため, 初期検索で得られたブログ記事から新たにトリガー対を発見することで, トリガーモデルをクエリに

動的に適応させる。以下に、提案するモデル適応の手順を示す。

- (1) 所与のクエリ q について、任意の検索モデルで初期検索を行う。
- (2) 上位 k 件のブログ記事を利用して、新しいトリガー対 $a \rightarrow b$ を発見する。また、これらのトリガー対について、関連度 $\alpha'(b|a)$ を算出する。
- (3) トリガーモデル $P_T(\cdot)$ を構築する。このとき、アマゾンの顧客レビューから得た関連度 $\alpha(b|a)$ または前ステップで得た $\alpha'(b|a)$ のうち、大きい方の値を用いる。

この適応により、所与のクエリに関する初期検索で上位に現れたブログ記事中で顕著なトリガー対を発見し、クエリにより特化した表現を含む主観的言語モデルを構築することが出来ると考えられる。

3. 評価実験

3.1 データ

提案手法の有効性を評価するため、TREC ブログトラック 2006 のテストコレクションを用いて評価実験を行なった。このコレクションは、2005 年の 12 月から 2006 年の 2 月までの 11 週間の間に収集された 3.2 百万のブログ記事から成り、50 個の情報要求（トピック）を持つ。図 1 にトピックの例を示す。

Topic #	851
Title	March of the Penguins
Desc.	Provide opinion of the film documentary "March of the Penguins".

図 1 TREC 2006 ブログトラックで用いられたトピックの例。

3.2 言語モデルの評価

本節では、まずアマゾンのレビューから構築された主観的言語モデルが、主観的なブログ記事を適切に表現することが出来るか否かをパープレキシティにより評価する。一般的に、低いパープレキシティを持つモデルほど良い（より正確な予測が行える）言語モデルだと言える。この尺度を用いて言語モデルの特性を調査するため、ブログトラックのテストコレクションのうち、意見を含むブログ記事を連結し、仮想的に長大な意見文書 d_O を作成した。同様に意見を含まないブログ記事全てを連結し、非意見文書 d_N を作成した。表 3 に、これら二つの仮想文書上でのいくつかの条件下でのパープレキシティ値を示す。

この結果から、三つの重要な事実が観測できる。まず、高次の n ほどパープレキシティが単調に低下し、より良く意見記事を表現出来ている。次に、意見文書 d_O は非意見文書 d_N よりも低いパープレキシティを示しており、アマゾンのレビューから構築した言語モ

表 3 パープレキシティの算出値。括弧内の数字は対応する P_B と比較したときのパープレキシティの減少率。

n	非意見 (d_N)		意見 (d_O)	
	P_B	P_E	P_B	P_E
1 グラム	9369	8946 (-4.5%)	7198	6829 (-5.1%)
2 グラム	6526	6279 (-3.8%)	4749	4546 (-4.3%)
3 グラム	5998	5762 (-3.9%)	4337	4145 (-4.4%)

デルでも、意見を含むブログ記事の特徴を反映出来ていることが分かる。最後に、主観的言語モデル P_E はベースライン言語モデル P_B よりも低いパープレキシティを示しており、通常の n グラム言語モデルでは捉えられない特徴を主観的トリガー対が捉えていることが分かる。

3.3 主観的言語モデルを用いた意見情報検索

3.3.1 初期検索

初期検索には、言語モデルを用いた推論ネットワークモデル⁶⁾ (INM) を採用した。索引付けの際は禁止語を除去し、大・小文字の区別および接辞の除去は行っていない。表 4 に、初期検索の結果と先行研究の結果を示す。なお、先行研究の結果は、意見情報検索のための様々な機能を用いて得られた結果であるのに対し、INM は単なる初期検索の結果であることに注意する必要がある。

表 4 初期検索の結果と TREC (以降) の結果との比較。

		MAP
初期検索	INM	0.1965
TREC	Best	0.1885
	Median	0.1156
TREC 以降	W. Zhang ら ²⁾	0.2726
	M. Zhang と Ye ⁷⁾	0.2257

3.3.2 主観的言語モデルと検索モデルの統合

INM による初期検索では、それぞれのブログ記事 d に関して、 d がユーザのクエリ q に適合する確率 $P(q|d)$ が推定されている。所与の d が主観的意見を含むか否かは d が q に適合するか否かと独立であると仮定すると、 d が適合記事でありかつ意見を含む確率は、 $P(q|d)$ と $P_E(d)$ $\approx \prod_{i=1}^m P_E(w_i|h_i)$ の積で表せる。しかしながら、長い文書はより小さい $P_E(d)$ を持つ傾向にあり、また二つの確率分布は全く異なる分散を持つ可能性がある。そこで両者の重み付き対数和を取り、かつ $P_E(d)$ を文書の長さ m で正規化することで、初期検索の再順位付けに用いる最終的なスコア $Scr(d, q)$ を得る。

$$Scr(d, q) = (1 - \beta) \log P(q|d) + \frac{\beta}{m} \sum_{i=1}^m \log P_E(w_i|h_i) \quad (3)$$

ここで β は主観的言語モデルの影響を制御するパラメタである。

式(3)の β を0から1まで徐々に増加させながら、初期検索の結果を再順位付けしたときのMAP値の変化を図2に示す。ここで、一番左($\beta=0$)の結果が初期検索に対応する。 β を変化させることで、MAP値は顕著に向上し、 $\beta=0.35$ のとき0.2398(+22.0%)の性能が得られた。

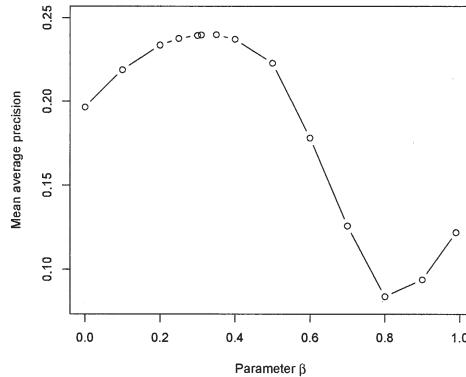


図2 意見情報検索におけるパラメタ β とMAP値の関係。

さらに、個々のクエリに関して主観的言語モデルの影響を分析したところ、顕著な精度向上が見られたトピックの多くは「製品」であった。本研究で主観的言語モデルの構築に用いたアマゾンのレビューが対象にしていない製品、たとえばビールや薬品に関しても精度向上が見られ、これらのレビューから獲得した言語モデルが製品全般、およびある種の非製品に関しても一般化可能であることを示唆する結果となった。統いて、システム性能が少しでも低下したトピックに関して調査したところ、特に「政治」と「組織」に関連するトピックで検索精度が低下する場合があった。すなわち、政治や組織等に関する主観的意見を表現するために用いられる語彙は、製品に関するそれとは異なるものと考えられる。

3.3.3 モデル適応

表5に、前出の結果とモデル適応を行った場合の結果との比較を示す。モデル適応の際は、1) クエリ(トピックのtitle)のみ、2) 14の代名詞のみ、3) クエリと代名詞の両方をトリガー候補として用いた。表中のアスタリスクは、 $p < 0.01$ の有意な性能向上が見られた試行を示す。

トリガーモデルをクエリに適応することで、いずれの場合も若干の精度向上が見られた。個別のクエリについてさらに精査すると、最も顕著な性能向上が得られたのは「Zyrtec」(薬品)であり、平均精度が0.2187から0.3230(+47.7%)へと向上していた。

次に、前節で議論した難易度の高いトピックに関して、モデル適応の効果を個別に観察したところ、効果

表5 初期検索と主観的言語モデルによる再順位付け後の結果とモデル適応後の結果との比較。

条件	MAP	初期検索からの向上
初期検索	0.1965	—
再順位付け後	0.2398	22.0%
1) クエリのみ	0.2430	23.6%
モデル適応 2) 代名詞のみ	0.2456*	25.0%
3) クエリ+代名詞	0.2452*	24.8%

は限定的ではあるものの、モデル適応によって多くのトピックで肯定的な影響が見られた。

4. おわりに

本稿では、トリガー言語モデルの意見情報検索への適用を議論した。TREC プログラム 2006 のデータを用いた実験では、提案モデルを利用することで、MAP 値で 22.0% の検索精度向上が見られた。さらにクエリへのモデルの動的な適応を行ったところ、通常の主観的言語モデルでは扱いにくいクエリに関しても若干の精度向上が得られた。著者らが知る限り、これまで言語モデルを用いて主観的意見に特徴的な遠距離にある語間の依存関係を捉え、かつ意見情報検索に効果的に適用した例はない。

参考文献

- 1) Mishne, G.: Multiple Ranking Strategies for Opinion Retrieval in Blogs, *Proc of the 15th TREC* (2006).
- 2) Zhang, W., Yu, C. and Meng, W.: Opinion retrieval from blogs, *Proc of the 16th ACM CIKM*, pp.831–840 (2007).
- 3) Zhou, G., Joshi, H. and Bayrak, C.: Topic Categorization for Relevancy and Opinion Detection, *Proc of the 16th TREC* (2007).
- 4) Lau, R., Rosenfeld, R. and Roukos, S.: Trigger-based language models: a maximum entropy approach, *Proc of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol.2, pp.45–48 (1993).
- 5) Tillmann, C. and Ney, H.: *Grammatical Interference: Learning Syntax from Sentences*, Lecture Notes in Computer Science, chapter Selection criteria for word trigger pairs in language modeling, pp.95–106 (1996).
- 6) Metzler, D. and Croft, W.: Combining the Language Model and Inference Network Approaches to Retrieval, *IP&M*, Vol.40, No.5, pp. 735–750 (2004).
- 7) Zhang, M. and Ye, X.: A generation model to unify topic relevance and lexicon-based sentiment for opinion retrieval, *Proc of the 31st ACM SIGIR*, pp.411–418 (2008).