

# Yahoo!ニュースにおける建設的コメント順位付けモデルの導入

田淵 義宗<sup>1</sup> 小林 隼人<sup>1,2</sup> 村尾 一真<sup>1</sup>

<sup>1</sup> ヤフー株式会社 <sup>2</sup> 理化学研究所 AIP センター

{yotabuch, hakobaya, kmurao}@yahoo-corp.jp

## 1 はじめに

オンライン・ニュースサービスにおいて、ニュース記事に投稿されるユーザのコメントはユーザ生成コンテンツ (*user-generated content*; *UGC*) とも呼ばれる有益なコンテンツのひとつである。ユーザは記事と共にこれらのコメントを読むことで、記事に関する補足情報を得ることができる。しかし、1 ページに表示できるコメント数は限られており、ユーザはすべてのコメントに目を通すことができないため、何らかの方法でコメントを順位付けする必要がある。どのようなコメントを優先的に表示するかはユーザ満足度に直結するため、この順位付けの改善はサービスにとって重要な課題である。

オンライン・ニュースサイトやオンライン・フォーラムにおけるコメントの順位付けに関する研究は既にいくつか存在するが [1, 2, 3, 4], これらの研究は基本的に評価用ボタンに基づくユーザ評価値のみを用いて順位付けを行っている。評価用ボタンの実例としては、図 1 のように各コメントに「そう思う/そう思わない」といった 2 択のボタンを表示しているものなどがある。これらはユーザの共感を表すためには効果的だが、この評価値のみで順位付けをすることには 2 つ問題がある。1 つ目の問題は、ユーザ評価値が多数派の意見に影響されやすく、必ずしもサービス提供側のニーズを満たさないということである。例えば、男性ユーザが多数を占める場合にユーザ評価値で順位付けを行うと男性の好むコメントを上位に出し続けてしまうが、これでは女性ユーザを切り捨てることになるため長期的には良い戦略とは言えない。2 つ目の問題は、ユーザ評価値が位置バイアスを受けやすいということである。多くのユーザは順位付けされたコメントのうち上位の少数コメントしか読まないため、時間が経って投稿されたコメントのほとんどが評価されることなく無視されてしまい、質の高いコメントが埋もれてしまう。

そこで本論文では、直接コメントの良さを推定するために「建設的度合い」を利用し、コメントの順位付



図 1: Yahoo!ニュースにおけるコメントの表示例 (元記事は大谷翔平選手の大リーグ新人王受賞に関する)。

けを行う方法について検討する。建設的度合いについては過去に研究がなされているが [5, 6], 順位付けに応用したものは存在しなかった。これまで我々は、コメントが建設的かどうかの 2 値ラベルではなく、建設的度合いを表す (0 から 10 の) 数値ラベルのついたデータセットを構築してきた [7]。本論文では、このデータセットを用いて構築した順位付けモデルを Yahoo! ニュースで実用化したときの具体的な取り組みについて報告する。

以降の構成は以下の通りである。2 節で「建設的」の定義と、構築したデータセットについて説明する。3 節で順位付けモデルを学習するためのペアワイズ法と、後処理でのスコアの正規化について説明する。4 節で特徴量選択のための実験、並びに、編集者による定性評価と現行モデルとの A/B テストの結果について説明する。5 節でまとめと今後の課題を述べる。

## 2 データ

### 2.1 「建設的」の定義

先行研究 [7] と同様に、Kolhatkar と Taboada の「建設的」の定義 [6] を参考に、表 1 に示す前提条件と主条件からなる定義を採用した。前提条件は最低限のコメントの質を担保する条件で、主条件は建設的となりうる状況を示す条件である。建設的コメントは、前提

前提条件	●記事に関しており、誹謗中傷を含まない
主条件	●自分の意見を元に議論を引き起こそうとする発言 ●客観的で、必要であれば根拠を提示している発言 ●新たな考え方、解決策、洞察を提供する発言 ●記事に関する珍しい経験談

表 1: 建設的コメントの定義（建設的コメントは前提条件を満たし、主条件の少なくとも1つを満たす）。

条件を満たし、主条件の項目を1つ以上満たすコメントとして定義される。

## 2.2 データ詳細

先行研究 [7] で作成されたデータセットの一部を用いて実験を行う。このデータセットはクラウドソーシングにより収集されたもので、各コメントについて10人の作業者が建設的かどうかを判定し、その結果の投票数(0~10)が建設的度合いとして付与されている。本研究では、49,215件のコメント(記事数9,843件で各5件のコメント)を、訓練セット80%、確認セット10%、テストセット10%に分けて用いた。

## 3 手法

### 3.1 ペアワイズ法

本研究では順位付けモデルの学習でよく使われるペアワイズ法を用いる。この方法では、2つのコメントが与えられたとき、それらの大小関係の2値分類問題を解くことで順位付けモデルを構築する。ここでは、線形モデルを用いて次のように最適化を行う。まず2節で説明したデータ  $D$  を、記事  $q$ 、コメント  $x$ 、建設的度合い  $y$  の3つ組  $(x, y, q) \in D$  で表されるものとする。このとき、 $q_i = q_j$  となる同じ記事における  $y_i > y_j$  となる順位付けられたコメントのIDペア集合  $P = \{(i, j) \mid q_i = q_j, y_i > y_j, (x_i, y_i, q_i) \in D, (x_j, y_j, q_j) \in D\}$  について、次式で定義される最適化問題を解くことで線形モデルの順位付け学習が実現される。

$$\min_w \frac{1}{2} w^\top w + C \sum_{(i,j) \in P} \ell(w^\top \tilde{x}_i - w^\top \tilde{x}_j). \quad (1)$$

ここで、 $w$  は学習される重みベクトル、 $\tilde{x}$  はコメント  $x$  から抽出した特徴量、 $C$  は正則化パラメータである。 $\ell$  は二乗ヒンジ損失であり、 $\ell(d) = \max(0, 1 - d)^2$  で定義される。

### 3.2 スコア正規化

4.2節で後述するが、今回定義した建設的要件はその性質上コメントが長いほど建設的だと判定されやすい。この現象は定義としては正しいが、実用上は読み

やすさを考慮して長さで正規化を行いたい。そこで本研究では、モデルが出力したスコアを後処理で調整するヒューリスティックな方法を採用した。具体的には、 $f(x) = w\tilde{x}$  を順位付けスコアとしたとき、次式で表される調整値  $f^*(x)$  を代わりに用いて順位付けを行う。

$$f^*(x) = f(x) - \alpha|x|. \quad (2)$$

この式は、順位付けスコアの期待値がコメントの長さに対して線形に増加すると仮定したとき ( $E[f(x) \mid |x| = L] = \alpha L + \beta$ ) に、その期待値がコメントの長さ依存せず一定となるような変換によって得られたものである。実際のコメントの分布は必ずしも仮定を満たすわけではないが、式の意味を考えると、短くても建設的なコメントが出やすくなるのが期待される。実験では順位付けスコアを線形近似し、その増加率を  $\alpha$  として設定した。

## 4 実験

### 4.1 特徴量選択

3節で説明した順位付けモデルについて、特徴量を変えた下記4つのモデルを用意し比較を行った。

- Uniq: コメント内の語彙数 (1次元ベクトル)
- TfIdf: コメント内の単語の Tf-Idf ベクトル
- OneHot: コメント内の単語の One-hot ベクトル
- Tf: コメント内の単語の出現頻度ベクトル

前処理としては、内製の解析器による形態素解析のみを行った。学習時の正則化パラメータ  $C$  は、 $\{2^1, \dots, 2^{13}\}$  から開発セットで選択した。

表2に平均 NDCG (normalized discounted cumulative gain) とペアワイズ精度の結果を示す。NDCG は順位付けモデルの性能評価をする際に一般的に用いられている指標で、 $NDCG_k = Z_k \sum_{i=1}^k \frac{2^{r_i} - 1}{\log(i+1)}$  で計算される。ここで、 $Z_k$  は最大値を1にする正規化係数、 $r_i$  はモデルが  $i$  番目に順位付けした要素の関連度スコア (ここでは建設的度合い) である。添字  $k$  は上位何件を評価するかを表しており、本研究では  $k$  についての平均値  $\sum_{k=1}^5 NDCG_k$  を最終的な評価指標とした。ペアワイズ精度はテストセット中のコメントの大小関係の正解率である。

表からは、NDCG、ペアワイズ精度ともに Tf が高い性能を示していることが分かる。OneHot は Tf と似た特徴量であるため、性能も同程度であった。TfIdf については、期待に反して高い性能が得られなかった。これは、今回の順位付け問題については機能語に有益

	平均 NDCG	ペアワイズ精度
Uniq	0.757	73.24%
TfIdf	0.739	70.55%
OneHot	0.813	77.41%
Tf	<b>0.818</b>	<b>77.89%</b>

表 2: 特徴量を変えたときの順位付けモデルの平均 NDCG とペアワイズ精度.

な情報が含まれているからだと考えられる. 例えば, 「です」「ます」などの機能語は, 記事のカテゴリ分類の性能には寄与しないが, 建設的度合いの予測には役立ちうる. Uniq はコメントの長さを表す単純な特徴量ではあるが, TfIdf よりも高い性能を示した. ただし, 次節の編集者による評価ではコメントの長さによる順位付けはあまり良い結果が得られなかった. 以上により, 本論文では一番性能が良かった Tf を採用した.

## 4.2 編集者評価

現行モデルに対する順位付けモデル (Tf 特徴量) の効果を検証するために, 編集者による定性評価を実施した. 現行モデルは, コメントの誹謗中傷などに関する予測値を総合して計算されたスコアで順位付けを行うモデルである. この現行モデルのスコアに建設的度合いを追加したモデルを建設的順モデルと呼び, 比較を行う. 評価は複数の編集者による相対評価の平均値で行った. 具体的には, 各記事について評価対象の複数のモデルにより順位付けされたコメントリストを 5 名の編集者が順位付けし, そのマイクロ平均 (平均順位) を評価値とした. 編集者には, 「どのリストが建設的か」ではなく, 「サービスとしてどのリストを提供すべきか」という観点からの評価を依頼している.

表 3 に, 編集者評価の結果を示す. 現行モデル, 建設的順モデルの他, 参考のため新着順, 長さ順の順位付けモデルとの比較を合わせて行った. 結果を見ると, 建設的順モデルが最も平均順位が良く, 編集者に高く評価されていることが分かる. 長さ順モデルは単純な方法ではあるが, 現行モデルよりは建設的なコメントを提示出来ている. ただし, 長さ順モデルは内容を考慮していないため, 建設的順モデルには大きく引き離される結果となった. 新着順はほぼランダムな並びとなるため, 予想通り現行モデルよりも悪い結果となっている.

上記の結果から建設的順モデルが現行モデルよりも良いコメントリストを提供できることが分かったが, 評価を行った複数の編集者から「建設的順モデルのり

	現行	新着	長さ	建設的
平均順位	2.61	3.42	2.20	<b>1.77</b>

表 3: 編集者評価における現行, 新着順, 長さ順, 建設的順モデルの平均順位.

	現行	建設的	正規化	長さ制限
平均順位	3.36	2.35	<b>2.10</b>	2.18

表 4: 編集者評価における, 現行, 建設的順, 正規化, 長さ制限モデルの平均順位.

ストは長いコメントばかりだが, 短くても建設的なコメントであれば上位に表示したほうが良い」というフィードバックを受けた. これはコメントが長いほど建設的要件が満たされやすくなることに起因しており, この定義の下では正しい事象ではあるが, 実用上は何らかの対処が必要である. そこで, 本研究では後処理で順位付けスコアを調整する方針を選択した. 詳細は 3.2 節に譲るが, 順位付けスコアがコメントの長さに依存しないように正規化を行うモデルを別途用意し, 再度比較を行った.

表 4 に, 2 回目の編集者評価の結果を示す. 参考のため, 単純に閾値以上の長さのコメントのスコアを 0 にするモデル (長さ制限モデル) との比較も行った. 閾値は, 編集者にアンケートをして 200 文字に設定した. 結果からは, 正規化モデルが最も良いコメントリストを提供できていることが分かる. 現行・建設的順モデルともに表 3 とは異なる数値となっているが, これは平均順位が他のモデルとの相対評価であるため自然な結果である. 長さ制限モデルについては, 建設的順モデルよりも高い評価を受けているが, 正規化モデルに及ばない結果となった. 実用的観点からも, 長いコメントでも建設的であれば表示すべきであるため, 総合的に見て正規化モデルの方がサービス導入に適していると考えられる.

表 5 は現行, 建設的順, 正規化モデルによって選ばれる先頭コメントの典型的な具体例である. この例では, 現行モデルが無難な短いコメントを選んでいるのに対して, 建設的順, 正規化モデルは内容のあるコメントを選んでいることが分かる. また, 正規化モデルは建設的順モデルよりも短いコメントを選ぶことが出来ている.

## 4.3 A/B テスト

現行モデル, 建設的順モデル, 正規化モデルについて, Yahoo!ニュースのコメント欄に実装して A/B テスト

(現行)
4 割はこの業者を使うか気になる
(建設的)
ヤマトはペリカン便を吸収して大きくなった経緯がありますよね。 忙しい、人手が足りないから即値上げてどうなんでしょうか。ガソリン代が高騰したとか、社員のお給料を確実に上げるとか、明確な理由があっての「値上げ」なら納得するのですが... 当方が定期コースで購入している某企業は宅配業者をヤマトからゆうパックに変更しました。 時間帯指定がヤマト運輸より緩やかですが、特に不便は感じません。 宅配ボックスのあるマンションや企業によってはコンビニ受け取りができるところもありますよね。 どうもヤマト運輸が顧客企業の足下を見てるような感じの値上げに思ってしまう=そのツケはカスタマーの我々が支払うことになるので余計にそう感じるのかもしれませんが。 上は皆さんのご意見に反して厳しいかと思いますが、3社の寡占状態だから値上げが行われるのであれば、中小の運送業者さんにも頑張ってもらいたいです。
(正規化)
今まで運送会社の経営陣は現場の方々を支払うべき対価を支払わずに利益を得ていたのに、今回の値上げを消費者に背負わせるだけのつもりだろうか？今までの経営陣や株主が得てきた利益を現場に還元してから値上げを了承してもらおう立場にいるのではないかと思う。経営陣や株主らは自身の身を切らずに取引先や消費者への負担ばかりをお願いしているようにしか感じられない。何も運輸だけでは無い問題だと思う。所謂消費に使え分の金額が給与として反映されていないのに、株価などだけで景気が良いと訴えている限り、現場は潤わないし、消費も上向くわけが無い様に思えるのだが、果たして。

表 5: 現行, 建設的順, 正規化モデルの先頭コメントの具体例 (「運送業者がネットショッピング事業者に宅配便の値上げを求め, 6割が応じた」という記事)。

を実施した。評価指標として下記 3 つの数値を算出した。

- リスト CTR: 記事詳細 (コメント 3 件のみ表示) のコメントリスト表示リンクのクリック率<sup>1</sup>。
- 既読 Imps: コメントモジュールの中央部が画面に 1 秒以上入った回数。
- 返信 CTR: 各コメントの返信コメント追加表示リンクのクリック率。

表 6 に 2 週間実施した A/B テストの結果を示す。数値はすべて現行モデルを 100 としたときの割合である。リスト CTR を見ると, 建設的順モデル, 正規化モデルともに数値が減少していることが分かる。これは, 建設的なコメントは長くなる傾向があるため, ユーザが記事詳細の 3 件のコメントに満足して離脱しているからだと考えられる。既読 Imps についても同様の傾向が見られるが, 建設的な (長い) コメントの方が読み飛ばされる確率が下がるため離脱の影響を打ち消している。返信 CTR は逆に, 建設的なコメントの方が高い数値となっている。この結果は, 返信コメ

<sup>1</sup>記事詳細ではコメント 3 件のみが表示され, ユーザが他のコメントを読みたい場合はコメントリスト表示リンクをクリックする。

	リスト CTR	既読 Imps	返信 CTR
現行	100.0	100.0	100.0
建設的	97.0	99.4	111.6
正規化	99.0	99.9	107.2

表 6: Yahoo!ニュースのコメント欄における現行, 建設的順, 正規化モデルの A/B テストの結果。

ントを読みたくなるような建設的なコメントが表示されていることを示唆している。以上の結果より, リスト CTR, 既読 Imps とともに現行モデルと同程度で, 編集者評価で良好な結果を示した正規化モデルを本採用とした。

## 5 おわりに

本論文では, ニュース記事における建設的コメントの順位付けモデルを Yahoo!ニュースに導入する取り組みについて報告した。今後の課題としては, 今回の順位付けモデルが出力する静的なスコアとリアルタイムで変化する動的なユーザ評価値とを上手く調整する方法を検討することなどが挙げられる。

## 謝辞

編集者評価と A/B テストについては, Yahoo!ニュースに関わる多くのエンジニア・編集者の協力を受けて実施された。ここに感謝の意を表する。

## 参考文献

- [1] Chiao-Fang Hsu, Elham Khabiri, and James Caverlee. Ranking Comments on the Social Web. In *Proceedings of CSE 2009*, Vol. 4, pp. 90–97. IEEE, 2009.
- [2] Anish Das Sarma, Atish Das Sarma, Sreenivas Gollapudi, and Rina Panigrahy. Ranking Mechanisms in Twitter-like Forums. In *Proceedings of WSDM 2010*, pp. 21–30. ACM, 2010.
- [3] Dirk Brand and Brink Van Der Merwe. Comment Classification for an Online News Domain. In *Proceedings of UMICTA 2014*, pp. 50–55. Stellenbosch University, 2014.
- [4] Zhongyu Wei, Yang Liu, and Yi Li. Is This Post Persuasive? Ranking Argumentative Comments in Online Forum. In *Proceedings of ACL 2016*, pp. 195–200. ACL, 2016.
- [5] Courtney Napoles, Aasish Pappu, and Joel R Tetreault. Automatically Identifying Good Conversations Online (Yes, They Do Exist!). In *Proceedings of ICWSM 2017*, pp. 628–631. AAAI Press, 2017.
- [6] Varada Kolhatkar and Maite Taboada. Constructive Language in News Comments. In *Proceedings of the First Workshop on Abusive Language Online*, pp. 11–17. ACL, 2017.
- [7] 藤田総一郎, 小林隼人, 奥村学. 建設的ニュースコメントの順位付けのためのデータセット構築. NL 研, pp. 2018–NL-236(14), 2018.