The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

疑似順位付け評価を用いたニュースコメント順位付けモデルの教師なしアンサンブル

Unsupervised Ensemble of Ranking Models for News Comments Using Pseudo Ranking Evaluation

藤田 綜一郎 *1 Soichiro Fujita　小林 隼人 *2*3 Hayato Kobayashi　奥村 学 *1 Manabu Okumura

*1東京工業大学 Tokyo Institute of Technology　*2ヤフー株式会社 Yahoo Japan Corporation　*3理研 AIP RIKEN AIP

In this paper, we propose a simple unsupervised ensemble method that determines the importance of each model by comparison of multiple models like majority vote. Experimental results on a news comment ranking task show that our proposed method outperforms current ensemble methods including the supervised one.

1. はじめに

オンラインニュースサイトには，ニュース記事毎に読者の議論の場としてコメント欄が設けられているものがある．それらのサイトの多くは，肯定的な読者評価が多く寄せられた順にコメントをランキング（順位付け）し，読者に良いコメントを優先的に提示している．しかし，読者評価の数は単調増加するため，単純に早期に投稿されたコメントほど多くの評価を得やすく，コメントの良さが平等に評価されているとは言い難い．そのため，藤田ら [藤田 18] はコメントの建設的度合いに注目し，良いコメントを直接ランキングする事に取り組んだが，分類器の精度が高くないことが課題として残されていた．

一方で，分類器のアンサンブルは機械学習モデルの精度を向上させる手法として広く知られている．最近では，Kobayashi [Kobayashi 18] が要約タスクにおいて，モデルの出力間のコサイン類似度を用いてモデルの多数決をとることで高速に動作し，かつ高い精度が得られたという報告がある．我々は，藤田らが作成したデータセットが読者が建設的かどうかの多数決によりラベル付けされていることから，コメントのランキングにも Kobayashi の類似度を用いた多数決によるアンサンブルモデルが適していると考えた．

そこで，本研究では，ランキングの評価指標 NDCG@k を用いて，記事毎に対応したモデルの重みを求める，多数決ベースの教師なしアンサンブル手法を提案する．具体的には，学習した複数のモデルの出力の平均によるランキングを擬似正解とみなし，各モデルの NDCG@k を求める．NDCG@k の値を記事に対するモデルの重要度とみなし，重み付けやモデル選択を行う．

ニュースコメントの建設的度合いをランキングするタスクの実験を行い，結果として，モデルの出力を NDCG@k の値で重み付けすることで既存手法を上回る精度を達成した．また，NDCG@k が高いモデルだけを選択してアンサンブルすることで，さらなる精度向上が見られた．

2. 提案手法

2.1 LSTM ランキングモデル

本研究では，Burges ら [Burges 05] の RankNet をベースにし，記事のタイトル $q \in Q$ とコメント $c \in C$ の二つを入力とし，記事 q におけるコメント c のランキングスコアを出力とする，LSTM ランキングモデルを使用する．

記事のタイトルとコメントの入力単語系列をそれぞれ $X_q = [x_1^q, ..., x_{|q|}^q]$, $X_c = [x_1^c, ..., x_{|c|}^c]$ とおく．まず，記事タイトルとコメントの単語系列を符号化する．符号化器の隠れ状態 $\boldsymbol{h}$ は次式で計算される:

$$\boldsymbol{h_i^q} = encoder_{title}(\boldsymbol{h_{i-1}^q}, \boldsymbol{e_i^q}), \quad (1)$$

$$\boldsymbol{h_i^c} = encoder_{comment}(\boldsymbol{h_{i-1}^c}, \boldsymbol{e_i^c}). \quad (2)$$

ここで，e_i^q, e_i^c はそれぞれ記事タイトル中の単語 x_i^q とコメント中の単語 x_i^c の事前学習された単語分散表現である．また，$encoder_{title}$ と $encoder_{comment}$ はそれぞれ記事タイトルとコメントの単語系列の符号化器であり，本研究では一層の LSTM を用いた．このとき，コメント c のランキングスコア s^c は次のように表される:

$$s^c = \boldsymbol{W}([\boldsymbol{h}_{|q|}^q; \boldsymbol{h}_{|c|}^c]) + \boldsymbol{b}. \quad (3)$$

ここで，$\boldsymbol{W}$ は重みベクトル，$\boldsymbol{h}_{|q|}^q$ と $\boldsymbol{h}_{|c|}^c$ は符号化器の隠れ層の最終状態，$\boldsymbol{b}$ はバイアス項である．

学習の際にはスコアを直接学習するのではなく，2 つのコメント $\{c_A, c_B\} \in C$ の予測結果を比較し相対的に学習するペアワイズ学習を用いる．このとき，損失関数 $loss_{AB}$ は式 (4) の様に交差エントロピーを用いて計算される:

$$loss_{AB} = -\bar{P}_{AB} \log P_{AB} - (1 - \bar{P}_{AB}) \log(1 - P_{AB}) \quad (4)$$

$$P_{AB} = \frac{1}{1 + e^{-\sigma(s^{c_A} - s^{c_B})}}, \quad (5)$$

$$\bar{P}_{AB} = \frac{1}{2}(1 + S_{AB}). \quad (6)$$

ここで，$S_{AB} \in \{-1, 0, 1\}$ はコメント c_A と c_B の相対正解スコアであり，c_A の正解スコアが c_B の正解スコアより高い場合は 1，低い場合は -1，同一の場合は 0 となる．

2.2 Post-Eval アンサンブル

ここでは，記事ごとに多数決で代表モデルを選択する手法 `Post-Eval` について説明する．Kobayashi の研究では，要約タスクにおいてモデルの予測単語の系列同士をコサイン類似度で比較することでモデルの重要度を教師なしに導出している．`Post-Eval` ではモデルが予測したランキングを予測単語の系列と同等のベクトル系列とみなすことで，ランキングタスク

連絡先: *1{fujiso, oku}@lr.pi.titech.ac.jp
*2hakobaya@yahoo-corp.jp

に Kobayashi の手法を適用可能にする．学習された複数のモデル M の記事 q に対する予測ランキング（予測スコアのリスト）集合を R^q とする．この時，モデル $m_i \in M$ の予測ランキング $r_i \in R^q$ の重要度を式 (7) で定義する．

$$post_{r_i} = \frac{1}{|S|} \sum_{r_j \in R^q : r_j \neq r_i} f(r_j, r_i). \tag{7}$$

ここで，$f(x, x^*)$ は評価指標であり，通常 x にはモデルによる予測ランキング，x^* には正解ランキングを代入する．`Post-Eval` では，正解ランキングの代わりに，他のモデルの予測ランキングを用い，モデル間の類似度を計算する．アンサンブル後の最終ランキング $\bar{r}$ は，式 (8) に示す通り $post_{r_i}$ が最大値となる，言い換えるとランキングが最も密集しているところのモデルを選択する．

$$\bar{r} = \underset{r_i \in R^q}{\operatorname{argmax}}(post_{r_i}). \tag{8}$$

本研究では，評価指標 f として，Normalized Discounted Cumulative Gain (NDCG@k) [Burges 05] を用いる．NDCG@k はランキングの有効性の評価として，情報検索などのランキング問題で広く用いられている評価指標であり，ランキング結果上位 k 件において，正解のランキングとの近さを示している．NDCG@k は次式で表される:

$$\mathrm{NDCG@}k = Z_k \sum_{i=1}^{k} \frac{s_i}{\log(i+1)}. \tag{9}$$

ここで，s_i はモデルによって順位付けられた i 番目のコメントの正解スコアを示している．また，Z_k は正規化項であり，モデルの出力が正解と同じ順序で並べられているときに最大値 1 をとるように設定されている．

2.3 Weighted-Eval アンサンブル

2.2 節で評価指標によるモデルの取捨選択について述べたが，この手法ではモデルの重み付き平均をとるアンサンブル手法を上回る精度は達成できなかった (詳細は 3.3 節で述べる)．

そこで，モデルの取捨選択の代わりに，評価指標を用いて重み付き平均を行う `Weight-Eval` アンサンブル手法を提案する．`Weight-Eval` による最終ランキング $\bar{r}$ は，式 (10) で表される:

$$\bar{r} = \sum_{r_i \in R^q} f(r_i, t^q) \cdot r_i, \tag{10}$$

$$t^q = \frac{1}{|S|} \sum_{r_i \in R^q} \frac{r_i}{||r_i||}. \tag{11}$$

ここで，式 (11) の t^q は記事 q の擬似正解ランキングである．擬似正解ランキングは，各モデルのランク系列 r_i を L2 正則化したものの平均で表される．`Post-Eval` では重要度算出の際にモデル間の類似度を全て計算していたが，擬似正解ランキングにより一括で重要度を求めることで，ランキングの系列を高速に比較できる．

また，モデルの重み付き和を求める前に k 個のモデルを取捨選択する `Select@`k を導入することで，`Weight-Eval` の更なる精度向上を図る．具体的には，モデルの重要度 $f(r_i, t^q)$ が大きい順に選んだ k 個のモデルだけを対象に重み付き和を取ることで，`Weight-Eval` と `Post-Eval` 双方の利点を得ることを目指す．

3. 実験

3.1 実験設定

データセット: 我々は，データセットとして，藤田ら [藤田 18] が報告している Yahoo!ニュース記事のコメントの建設的コメントランキングデータセットを用いた．データセットは，(記事タイトル, コメント, 建設的スコア) の 3 つ組からなる．建設的スコアは，40 人に対して各コメントを提示し，コメントが建設的である/建設的でないの二択質問をした際に，建設的であると判別した人数で定義されており，0 ～ 40 の整数値をとる．このスコアは，出来るだけ多くの読者を十分に満足させることを目的としており，コメントが建設的であると考える潜在的な人数を模擬する尺度となっている．本研究では，データセットのうち 1,300 記事 130,000 コメントを学習データ，113 記事 11,300 コメントを開発データ，200 記事 42,436 コメントをテストデータとして用いた．学習データと開発データは各記事からランダムに抽出された 100 コメントにラベル付けがされており，テストデータは実際のサービスを想定して各記事の全てのコメントにラベル付けがされている．

前処理: データの前処理として，形態素解析器 MeCab[*1] [Kudo 04] を NEologd 辞書 [*2] [Toshinori 17] と共に用いて単語分割を行った．また，数字を特殊なトークンに置き換え，片仮名は平仮名にすることで文字種の統一をした．機能語はランキングの精度に良い影響を与えているため，ストップワードは除去しなかった．各データセットで出現回数が 3 回未満の単語は低頻度語として切り捨てた．

モデルの学習: 単語の分散表現は，word2vec [Mikolov 13] により約 150 万件のラベル無しコメントを事前学習したものを使用した．分散表現の次元数は 300 として窓幅 5 単語の Skip-gram モデルを学習した．ランキングモデルのパラメータの学習には Adam (学習率: 1.0×10^{-4}) を使用し，ミニバッチサイズは 10，イテレーション数は 1.0×10^4 とした．各ミニバッチは，同一記事からランダムに 10 ペアを選択することで作成した．記事タイトルとコメントのエンコーダの隠れ状態の次元数はどちらも 300 とした．実験では，アンサンブルのために，上記設定でモデルの初期状態をランダムに変化させた 100 モデルを学習した．

評価指標: 式 (9) で説明した NDCG@k に加えて，2 つ目の評価指標として Precision@k を用いる．これは推定された上位 k 件のコメントが、正解の上位 k 件のコメントに含まれている割合として定義されている．実験では $k \in \{1, 5, 10\}$ の場合の評価を行った．なお，数値関連度を用いたランキングの設定においては Precision@k よりも NDCG@k の方が優れた指標であると言われている．

3.2 比較手法

本研究では以下の手法について比較を行った．`RankSVM` と `LSTMRank` はそれぞれ既存手法と本研究で作成した単一モデルによるベースラインである．`ScoreAverage`，`RankAverage`，`Top-k-Average`，`Normalized` は一般的によく用いられる，教師なしで事後的にモデルを組み合わせるアンサンブルの手法である．`Supervised` は，教師ありデータによりモデルに重み付けをしてアンサンブルする手法である．

- `RankSVM`: 藤田らの研究 [藤田 18] で最も精度が高かった `RankSVM` モデル. 既存研究におけるベースラインである．

∗1 http://taku910.github.io/mecab/

∗2 https://github.com/neologd/mecab-ipadic-neologd

	NDCG			Precision		
	@1	@5	@10	@1	@5	@10
RankSVM	73.38	74.59	76.01	15.5	30.20	38.95
LSTMRank	76.35	77.97	79.52	15.0	33.20	42.99
ScoreAverage	76.91	79.11	80.48	16.08	33.67	44.32
RankAverage	79.19	80.53	81.81	13.57	36.18	46.08
Top-k-Average	78.38	80.52	81.57	14.07	35.38	46.08
Normalized	79.83	80.77	82.16	**17.08**	37.18	46.48
Supervised	78.64	80.33	81.94	16.28	35.47	46.58
Post-Eval	77.18	80.09	81.24	14.57	35.58	45.78
Weight-Eval	**79.87**	81.39	82.17	**17.08**	**37.88**	46.63
+ Select@50	**79.87**	**81.43**	**82.33**	**17.08**	37.39	**47.34**

表 1: 建設的コメントのランキング実験における NDCG@k と Precision@k の結果 ($k \in \{1, 5, 10\}$).
すべての値はパーセンテージで表されている.

- LSTMRank: 2.1 節のランキングモデルを 100 モデル学習した中で最も高い精度が得られたモデル．アンサンブルではなく，単一モデルで予測する．
- ScoreAverage: コメントごとにモデルの出力スコアを平均し，その値が高い順に並べる．
- RankAverage: モデルごとにコメントのランキングを行い，その順位をコメントのスコアとする．全てのモデルのスコアを平均し，その値が高い順に並べる．
- Top-k-Average: Cormack らの手法 [Cormack 09]. 各モデルのランキング上位 k 件のみコメントを対象に，モデルの出力スコアの総和を計算し，その値が高い順に並べる．
- Normalized: Burges らの研究 [Burges 11] に代表される，各モデルの出力を正規化してから平均をとる手法．正規化にはさまざな方法があるが，本研究では各モデルの出力のランク系列をベクトルとみなし，L2 ノルムによる正規化を行う．
- Supervised: 各モデルの開発データセットでの NDCG@k をモデルの重要度とみなし，スコアの重み付き和が高い順に並べる．この手法では，モデルの重要度は記事に関係なく一定である．

3.3 実験結果

実験結果を表 1 に示す．まず，単一モデル同士を比較すると本研究で用いた LSTMRank は藤田らのベースライン RankSVM を上回る精度を達成した．Precision@1 のみ RankSVM を下回っているが，これは LSTMRank が最もスコアの高いコメントを見つけられなかった一方で，スコアが高いものを多く推定できていることを意味している．

また，アンサンブルの結果，全ての手法において，単一モデルの結果を上回る精度を達成した．特に提案手法である Weight-Eval は最高精度を達成しており，Select@50 により 100 モデルから重要度が高い順に 50 モデルを選択して重み付け和をとることで，更に NDCG@k が向上した．

Kobayashi の手法をランキングタスクに適用させた Post-Eval は LSTMRank よりも高い精度を達成した．これにより，評価指標を用いてモデル間で類似度を計算し，それを元に記事ごとにモデルを事後的に選択する手法が有効であることが確認できた．しかし，一方で，Normalized などの一般的なモデルの平均によるアンサンブルよりも低い精度となった．これは元々ランキングが相対的な比較によってなされているため，Post-Eval による確信度が高いモデルを選択するよりも，モデルの多様性を保持する方が性能向上に有効だからだと考えられる．

教師なし手法である Weight-Eval は教師ありでモデルの重要度を学習する Supervised よりも高い精度を達成した．よって，事前にモデルの重要度を教師データを用いて学習するより，ランキング間の類似度から事後的にモデルの重要度を決める方が良いことが確認できた．

4. 関連研究

ニュースのコメント欄をはじめとするオンラインフォーラムのコメントの分析は，近年多くの研究がなされてきた．中でもコメントのランキングについては，読者による評価を予測しランク付けする研究 [Hsu 09, Das Sarma 10] や，説得力のあるコメント順にランク付けする研究 [Wei 16] など様々な観点からランキングをする試みがされている．本研究に最も関連する建設的なコメント順にランク付けする研究 [藤田 18] では，ランキング用データセットを作成し，精度向上のためにタスクに適したデータ作成の方法について考察されている．我々は，モデルの精度向上の方面から更なる精度向上を図っている．

ランキングタスクのアンサンブルでは，多数決選択のように事後的にモデルを教師なしで選択する手法として，モデルの出力を正規化してアンサンブルしている手法 [Burges 11] や上位のコメントだけをフィルタリングしてからモデルの平均をとる手法 [Cormack 09] がある．また，情報検索において，クエリ依存のアンサンブルの重みを半教師ありで学習する研究 [Hoi 08] がある．彼らの研究は，記事毎にモデルの重みを求める我々の手法と発想は類似しているが，情報検索がクエリに対する一致度で順位付けをするタスクであるのに対し，コメントのランキングは記事 (クエリ) に類似していることが前提の上に順位付けするため，我々の試みとは大きく異なる，

評価指標 NDCG@k を元に分類器の精度を向上させる試みとして，NDCG@k を損失関数に組み込み直接学習する LambdaRank [Burges 07] やそれに決定木を組み合わせた LambdaMART [Burges 10] が提案されている．しかし，これらは事前に教師ありの学習を行なっており，教師なしで事後的にモデルを組み合わせる提案手法とは本質的に異なるため，本研究では比較を行っていない．

5. おわりに

本研究では，評価指標を用いて教師なしで事後的にモデルをアンサンブルする手法を提案した．実験により，提案手法が最も高い精度を達成し，モデル間のランキングの近さを評価指標により求め，その値を重要度としてモデルの重み付けや取捨選択に使用することで精度が向上することを示した．
今後の展望として，同一の構造の分類器ではなく，様々な種類の分類器を組み合わせによるアンサンブルを行いたいと考えている．また，提案手法と教師ありのアンサンブル手法との精度面/速度面での性能比較を検討している．

参考文献

[Burges 05] Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., and Hullender, G.: Learning to rank using gradient descent, in *Proceedings of the 22nd international conference on Machine learning*, pp. 89–96ACM (2005)

[Burges 07] Burges, C. J., Ragno, R., and Le, Q. V.: Learning to rank with nonsmooth cost functions, in *Advances in neural information processing systems*, pp. 193–200 (2007)

[Burges 10] Burges, C. J.: From ranknet to lambdarank to lambdamart: An overview, *Learning*, Vol. 11, No. 23-581, p. 81 (2010)

[Burges 11] Burges, C., Svore, K., Bennett, P., Pastusiak, A., and Wu, Q.: Learning to rank using an ensemble of lambda-gradient models, in *Proceedings of the Learning to Rank Challenge*, pp. 25–35 (2011)

[Cormack 09] Cormack, G. V., Clarke, C. L., and Buettcher, S.: Reciprocal rank fusion outperforms condorcet and individual rank learning methods, in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pp. 758–759ACM (2009)

[Das Sarma 10] Das Sarma, A., Das Sarma, A., Gollapudi, S., and Panigrahy, R.: Ranking Mechanisms in Twitter-like Forums, in *Proceedings of the Third ACM International Conference on Web Search and Data Mining (WSDM 2010)*, pp. 21–30, ACM (2010)

[Hoi 08] Hoi, S. C. and Jin, R.: Semi-supervised ensemble ranking (2008)

[Hsu 09] Hsu, C.-F., Khabiri, E., and Caverlee, J.: Ranking Comments on the Social Web, in *Proceedings of the 2009 International Conference on Computational Science and Engineering (CSE 2009)*, Vol. 4, pp. 90–97, IEEE (2009)

[Kobayashi 18] Kobayashi, H.: Frustratingly Easy Model Ensemble for Abstractive Summarization, in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4165–4176 (2018)

[Kudo 04] Kudo, T., Yamamoto, K., and Matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphological Analysis, in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, pp. 230–237, Association for Computational Linguistics (2004)

[Mikolov 13] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J.: Distributed representations of words and phrases and their compositionality, in *Advances in neural information processing systems*, pp. 3111–3119 (2013)

[Toshinori 17] Toshinori, T. H., Sato and Okumura, M.: Implementation of a word segmentation dictionary called mecab-ipadic-NEologd and study on how to use it effectively for information retrieval (in Japanese), in *Proceedings of the Twenty-three Annual Meeting of the Association for Natural Language Processing*, pp. NLP2017–B6–1, The Association for Natural Language Processing (2017)

[Wei 16] Wei, Z., Liu, Y., and Li, Y.: Is This Post Persuasive? Ranking Argumentative Comments in Online Forum, in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pp. 195–200, Association for Computational Linguistics (2016)

[藤田 18] 藤田綜一郎, 小林隼人, 奥村学 F 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, 研究報告自然言語処理 (NL), Vol. 2018, No. 14, pp. 1–7 (2018)