The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

ランク学習によるYahoo!知恵袋の見出し生成

Headline Generation of Yahoo! Chiebukuro Based on Learning to Rank

日暮立 † 小林隼人 †‡ 村尾 一真 † 増山 毅司 †
Tatsuru Higurashi Hayato Kobayashi Kazuma Murao Takeshi Masuyama

† ヤフー株式会社 ‡ 理化学研究所 AIP センター
Yahoo Japan Cooporation RIKEN AIP

In this paper, we propose a training method for headline generation based on learning-to-rank, so that the appropriate summarization of a question posted by the user is distributed in Yahoo! Chiebukuro. For headline generation we used data provided from National Institute of Informatics and generated supervised-data through crowdsourcing. As a result, effective results were obtained compared with conventional methods.

1. はじめに

Web 上で配信されているテキストコンテンツには, 新聞記事やコラムをはじめとするニュースサイト以外にも掲示板や SNS など, ユーザーが投稿したコンテンツ (UGC: User Generated Content) がある. UGC の一つに, 質問したいユーザーと回答したいユーザーをむすび, お互いに知恵や知識を Q&A 方式で共有できる Q&A サイトがある. Q&A サイトでは, ユーザーに回答可能な質問を見つけてもらうために投稿された質問に見出しを付けて提示することが多いが, 新聞記事とは異なり, 必ずしもユーザーが適切な見出しを付けているとは限らない. また, ユーザーの投稿時の負担を減らすために投稿フォームに明示的に見出しの項目を設けていない Q&A サイトも多く, その場合には質問文の先頭部分が見出しの代用として用いられている. そのため, このような Q&A サイトでは見出しが質問内容を適切に表していないことが原因で, 未回答の質問が回答可能なユーザに見つけてもらえないという課題がある.

日本で多くのユーザーに利用されている Q&A サイトの一つに Yahoo!知恵袋 [*1] がある. 図 1 に, Yahoo!知恵袋の見出しと投稿質問文の例を示す. Yahoo!知恵袋では, ユーザーの入力の負荷が高いことから投稿時に見出しを入力させていないため, 前述のとおり投稿質問文の先頭から一部を切り出して見出しとして提供している. 投稿された質問文は, Web または Yahoo!知恵袋が提供しているスマートフォンアプリ [*2] を通じて配信される.

先頭切り出しの見出しが質問内容を適切に表していない場合があるという問題は, 特にスマートフォンアプリのように表示領域が限られている場合に顕著である. 図 2 に, 図 1 の内容をスマートフォンアプリの通知機能を利用して配信した場合の見出しの表示例を示す. 図 1 では, 投稿者の意図としては片思いの初期段階の恋愛相談をしているにもかかわらず, 図 2 のように先頭から部分的に切り出した見出しではユーザーに質問内容を適切に伝えることができていないことが分かる. このようにコンテンツの内容を正しくユーザーに伝えられていないことは, ユーザビリティ低下に繋がり, 利用者数に悪影響を及ぼす可能性がある. 逆に, 投稿質問文の内容を正しく反映した見出しを生成してユーザーにコンテンツを提供することができれば, ユーザーが見たい質問を容易に見つけられるようになり, 回答率の向上が期待できる.

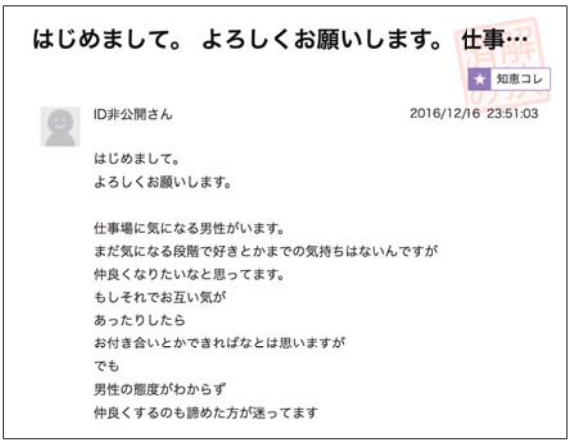

図 1: Yahoo!知恵袋の見出しと投稿質問文の例

図 2: Yahoo!知恵袋アプリの通知例

そこで本研究では, Yahoo!知恵袋の見出し生成タスクを投稿質問中の各文のランキング問題とみなし, ランク学習による見出し生成手法を提案する. その際, Yahoo!知恵袋のユーザーに最適な見出しを提供する目的で, クラウドソーシングを用いて各文の “見出しらしさ” のスコアを付与したデータセットを作成した. このデータセットを用いたランク学習モデルで生成し

連絡先: 日暮 立, ヤフー株式会社, thiguras@yahoo-corp.jp

∗1 http://chiebukuro.yahoo.co.jp/

∗2 https://itunes.apple.com/jp/app/id354626165

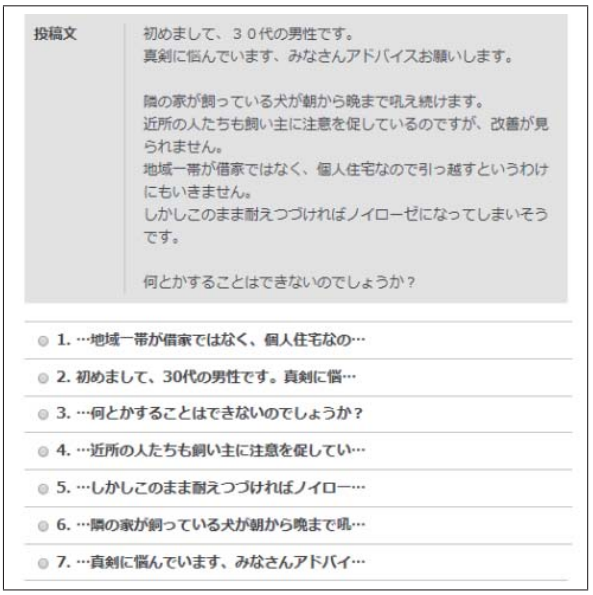

図 3: クラウドソーシングタスクの例

た見出しと, 現在 Yahoo!知恵袋で提供されている先頭切り出しや, 他の類似手法による見出しを比較評価したところ, 提案手法が最も有効であることが示された.

2. 関連研究

要約に関する研究は幅広くなされているが, その多くは新聞記事のようなある程度書式が統一されているデータに関するものであり, Q&A サイトのようにユーザーが自由に記述できるデータに対して要約技術を適用したものは少ない.

野口らは, 分散表現を用いて投稿質問文のスコアを定義し, ナップサック問題として定式化して動的計画法を解くことで Yahoo!知恵袋の要約を生成する教師なし学習の手法を提案し, 従来の TFIDF スコアを用いた抽出よりも良好な結果となることを示した [野口 15]. Bhaskar は, Q&A サイトにおいて, 複数文から成る質問文を, グラフを用いてクラスタリングした上でランキングし, ランクの最も高い文を要約文とする手法を提案した [Bhaskar 13]. Heilman と Smith は, 本文から質問文候補を文構造からルールベースで生成し, 質問文候補の教師あり学習によってランキングを学習する手法を提案している [Heilman 10]. しかし, これらの手法は質問文の要約タスクを対象としており, 本研究で扱っているような短い文字制約における見出し生成タスクについては検討していない.

我々のタスクに近い設定として, Ishigaki らは回答者の要旨理解のための長文質問の要約を生成的手法で行った [Ishigaki 17]. 彼らは, Yahoo! Answers のデータセットから投稿質問文を分析し, 投稿質問文と見出しを, 元文とその要約文の対とみなして教師データを抽出し, 生成型要約モデルを構築する方法を提案した. しかし, 生成型要約モデルは致命的な誤りを含む見出しを生成する可能性があるため, サービスに直接適用することは難しい.

3. 提案手法

本研究では, 見出しを持たない Q&A サイトを対象として, 教師あり学習を用いた抽出型の見出し生成手法を提案する. 教師あり学習を用いることでサービスに適した見出し生成モデルを構築することができ, 抽出型の手法を用いることでサービスの品質を維持することができる.

以下に, クラウドソーシングによる教師データの作成方法と, 作成したデータを用いた学習手法についてそれぞれ述べる.

3.1 データ作成

実験のためのデータ作成は Yahoo!クラウドソーシング *3 を利用して行った. データは国立情報学研究所がヤフー株式会社から提供を受けて研究者に提供している Yahoo!知恵袋のデータセット *4 を利用した. 以降, 知恵袋データと呼ぶ. 知恵袋データから投稿質問文を抽出し, 投稿質問文を文末表現 (「!」「?」「。」) で分割し, 分割した文の先頭 20 文字を見出し候補とした. 投稿質問文のうち, 上記の分割を行った場合に候補数が 6 以上となる投稿質問文 10,000 件を対象とした.

クラウドソーシングタスクの具体例を図 3 に示す. このタスクでは, 投稿質問文と, それに対する見出し候補となる選択肢が提示される. ワーカーには, 提示された投稿質問文を読んだ後, 他の投稿質問文と区別して内容が推測できそうな見出し候補を選択肢の中から一つ選ぶように指示をした. その際, 挨拶文, 自己紹介文や具体的な内容がない文を除くように補足説明も行った.

図 3 の例に対して, 各選択肢が見出しとして適切か不適切かを考える. 例題の投稿質問文から推測したい内容は「隣の家の犬の鳴き声に困っている」と考えられる. これを元に選択肢がそれぞれ適切かどうかの判断を行う. 選択肢 1, 選択肢 5 はペットに関する相談であることがわからないので不適切である. 選択肢 2, 選択肢 3, 選択肢 7 はユーザーへの挨拶文, 問いかけ文であり, 投稿質問文の内容に関する情報がないため不適切である. 選択肢 4 は「何かしらのペット」に関する内容であることが推測できるが, 具体的にどのようなペットなのか, またどのような相談なのかはわからないため, 不適切ではないが含まれる情報は不十分である. 選択肢 6 は 「隣の家の犬」に関する相談であることが推測できるため適切である. 以上の判断から選択肢 6 が最も適切な選択肢となる.

ワーカーには 1 タスクあたり, チェック設問（不適切に回答するワーカーを除外するための設問）を含めた 11 問を回答してもらった. 投票のポジションバイアスを除くために設問ごとに選択肢をランダムに並べ替えた. クラウドソーシングに掲載したタスク数は 10,000 タスクで, 1 タスクあたり 10 人のワーカーにタスクを依頼した. 1 質問あたり 10 人のワーカーが各々が最適と判断した選択肢を一つ選んでいるので, 各選択肢は 0 から 10 の投票スコアが与えられる.

3.2 学習方法

本研究では, クラウドソーシングで作成したデータを利用して, 投票スコアの高い選択肢の見出しを抽出する手法を提案する. クラウドソーシングで作成したデータの投票スコアの予測ではなく, 相対的に投票スコアの高い見出しを予測するという目的から, 本研究ではランク学習を利用する.

ランク学習は, 検索ランキングの最適化によく使われる手法であり, 検索クエリ q に対する文書 x とその適合度 y の 3 つ組の集合が与えられたときに, 検索結果を提示するための最適なランキングを学習することができる. 本研究においては, 投稿質問文を q, その見出し候補の選択肢を x, 各選択肢の投票スコアを y に対応付けて定式化を行う.

*3 http://crowdsourcing.yahoo.co.jp/
*4 http://www.nii.ac.jp/dsc/idr/yahoo/yahoo.html

本研究ではランク学習でよく使われる Pairwise 法を採用した. クラウドソーシングで作成したデータ D は, 投稿質問文 q, 選択肢 x, 各選択肢に対する投票スコア y の3つ組 $(x, y, q) \in D$ で表されるものとする. Pairwise 法では, データ D 中の3つ組のペア $((x_i, y_i, q_i), (x_j, y_j, q_j))$ について, $q_i = q_j$ となる同じ投稿質問中の, $y_i > y_j$ となる順位付けられた選択肢に基づいて学習を進める. 具体的には, まず上記を満たすようなデータ D 中の ID のペアを $P = \{(i, j) \mid q_i = q_j, y_i > y_j, (x_i, y_i, q_i) \in D, (x_j, y_j, q_j) \in D\}$ で定義する. この P を用いて, 下記で定義される最適化問題を解くことでランク学習が実現される.

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

ここで, w は学習される重みベクトル, $\tilde{x}_i$ は選択肢 x から抽出した特徴量, C は正規化パラメータである. ℓ は二乗ヒンジ損失であり, $\ell(d) = \max(0, 1 - d)^2$ で定義される.

4. 実験

4.1 実験設定

本実験では, クラウドソーシングで作成したデータで, 投稿質問文ごとの各選択肢に投票した数が多い順に順位をラベル付けした. クラウドソーシングで作成したデータ 10,000 件を学習セット 9,000 件と評価セット 1,000 件に分割した. 学習セット 9,000 件で投稿質問文ごとに各選択肢の順位を予測するランク学習を実行した. 学習では, 判定対象文とその前後の文に対して, Bag-of-Words (BoW) ベクトル, Doc2Vec ベクトル, 文長に関するベクトルを計算し, 9 つのベクトルを結合し特徴量とした. BoW ベクトルは, 文を形態素解析して抽出した名詞・動詞・感動詞・接続詞・副詞・形容詞の TFIDF スコアに基づく 30,820 次元のベクトルである. Doc2Vec ベクトルは, 文の意味の特徴を表したものであり, Yahoo!知恵袋のデータセットの投稿質問文 3,092,958 文を, gensim[*5] の doc2vec を用いて学習させたモデルで, 文を 100 次元のベクトルに変換した. 文長ベクトルは, 対象文の文字数をベクトルで表現したものである.

学習には RankSVM[*6] を使用した. 学習における正規化パラメータ C は, 交差確認でパラメータ探索を行い, 精度が最も高くなった $C = 0.125$ を採用した.

4.2 比較手法

実験では, ランク学習と以下の見出し生成手法との比較実験を行なった.

- **ランダム抽出**: クラウドソーシングの選択肢をランダムで一つ選択して見出しとした.
- **TFIDF 重要文抽出**: 各選択肢を形態素解析して抽出した名詞・動詞・感動詞・接続詞・副詞・形容詞の TFIDF スコアの合計が最も高い選択肢を見出しとして選択した.
- **TFIDF 類似文抽出**: 投稿質問文と各選択肢をそれぞれ形態素解析して抽出した名詞・動詞・感動詞・接続詞・副詞・形容詞の TFIDF スコアに基づく 31,990 次元のベクトル間のコサイン類似度を計算し, 投稿質問文と最も類似度が高い選択肢を見出しとして選択した.
- **Doc2Vec 類似文抽出**: gensim の doc2vec を利用して, 投稿質問文と各選択肢を 100 次元のベクトルに変換してコサイン類似度を計算し, 投稿質問文と最も類似度が高い選択肢を見出しとして選択した.
- **辞書ベース先頭削除**: Yahoo!知恵袋のデータセットにおいて, 出現頻度の高い挨拶文 (「おはようございます」「こんにちは」「はじめまして」など) や, 知恵袋の投稿質問文の先頭文として出現頻度の高い定型文 (「質問です」「教えて」「お願いします」など) を人手で 913 文抽出し, それらを投稿質問文から除外したうえで, 先頭 20 文字を切り出して見出し生成を行なった.

4.3 評価尺度

本研究では, 見出し生成のオフライン精度評価を行うための3つの指標を定義する. 以下では, 現行の Yahoo!知恵袋の投稿質問文の先頭切り出し見出しに対して, 各手法で生成した見出しが適切かどうかを評価する.

4.3.1 見出し変化率

先頭切り出しによる見出し以外の見出し候補（具体的には2文目以降から生成された見出し候補）が選ばれた場合, その見出し候補は各手法によって適切だと予測されたものだと判断できる. この予測が正しければ, 先頭切り出しによる見出しを変化させた割合が多い方が良い手法といえる. そこで, 見出し生成が行われて先頭切り出し見出しが変化した割合を評価指標のひとつとして, 以下で定義する.

$$\text{見出し変化率} = \frac{\text{先頭切り出し見出しが変化した質問数}}{\text{投稿質問文全て}} \tag{2}$$

4.3.2 勝率

先頭切り出しの見出しと比較して, 各手法が選択した見出し候補がクラウドソーシングの投票スコアの順位で上位であれば, その手法の予測は正しいとみなせる. そこで, 各手法による先頭切り出しに対する勝率を評価指標のひとつとして, 以下で定義する.

$$\text{勝率} = \frac{\text{生成見出しの投票スコアが最も高い質問数}}{\text{先頭切り出し見出しが変化した質問数}} \tag{3}$$

4.3.3 平均順位

生成された見出しが, クラウドソーシングの投票スコアで何位であるかの平均を各手法の最終的な評価指標として用いる. 平均順位は低いほど, より質の高い見出しを生成できたことを意味している.

5. 結果と考察

5.1 クラウドソーシングデータ分析

先頭切り出しの見出しの投票スコアが最も高い場合は見出しは変更しない方が良いと判断し, 先頭切り出し以外の選択肢の投票スコアが最も高い場合は, 見出しは変更した方が良いと判断する. この基準でクラウドソーシングの結果を分析したところ 38.1%の投稿質問文の見出しは, 先頭切り出しの見出しから別の見出しに変更した方が良いという結果になった. 従って, 全質問文に対して理想的に見出しが変更された場合の最適見出し変化率は 38.1%と言える. 先頭切り出しの見出しを選択した場合の平均順位は 1.582 であった.

クラウドソーシングの選択肢において, 挨拶文や自己紹介文が含まれる選択肢が選ばれることは少なかったが, 「教えてく

*5 https://radimrehurek.com/gensim/models/doc2vec.html
*6 https://www.csie.ntu.edu.tw/c̃jlin/libsvmtools/

表 1: 先頭切り出しと提案手法による見出しの生成例

	先頭切り出しによる見出し	提案手法による見出し
1	情緒不安定なものです。回答して下さった…	…新婚旅行は、行きは別々で行き、現地で…
2	カテ違いならゴメンナサイ。今、財布が破…	…今、財布が破れてツライ状況です。新し…
3	現在、私の父は 60 歳で定年退職を迎えま…	…厚生年金の支払いは何歳までなんですか…
4	前にとても大変で嫌な思いをしたので携帯…	…携帯メールお断りというのはこの落札者…
5	27 歳女です。環境的になかなか新しい出…	…環境的になかなか新しい出会いがなくて…

表 2: 見出し生成の評価結果

	見出し変化率 (%)	勝率 (%)	平均順位
先頭切り出し	0	-	1.582
ランダム抽出	85.9	11.7	2.784
TFIDF 重要文抽出	81.1	12.7	2.635
TFIDF 類似文抽出	79.3	18.7	2.396
Doc2Vec 類似文抽出	**88.2**	13.0	2.731
辞書ベース先頭削除	2.2	72.0	1.577
ランク学習	9.9	**94.9**	**1.569**

ださい」などの定型文を含む選択肢は 1 人または 2 人が選んでいることがあった. 説明文に記載している「具体的内容がない文」についてワーカーにより判断のばらつきがあったとみられる. クラウドソーシングにおいて, 手作業で全てのルールを網羅することは難しいので, ワーカーの主観により判断のばらつきが生じる場合があることを念頭におく必要がある.

5.2 定性評価

表 1 にランク学習によって見出しを生成した場合の例を示している. 生成した見出しの例 1, 例 2, 例 5 では先頭切り出しの見出しには「情緒不安定なものです」や「カテ違いならゴメンナサイ」や「27 歳女です」などの文が含まれていたが, 提案手法によって生成された見出しではそれらが除外されて表示されるようになり, より質問内容を推測しやすくなっている. また, 生成した見出しの例 3, 例 4 では, 先頭切り出しの見出しと比べて, 提案手法によって生成した見出しは, より内容の具体性が増し, 質問内容が推測しやすくなっている.

5.3 オフライン評価

オフライン評価として, ランク学習と比較手法で生成した見出しを, 4.3 節で定義した評価尺度を用いて比較した. 比較結果を表 2 に示す. ランダム抽出, 重要文抽出, 類似文抽出による手法では, 見出し変化率がクラウドソーシングのデータ分析から得られた最適見出し変化率 38.1%を大きく上回り, 適切な見出しを変化させて不適切な見出しに変化させていることがわかった. そのため勝率も 10%台という低い数値となっており, 平均順位も先頭切り出しの見出しを選んだ場合の 1.582 より悪い結果となっている.

一方で, 辞書ベース先頭削除による見出し生成の結果においては, 見出し変化率は 2.2%とカバレッジは高くないが, 先頭切り出しの見出しと比較して平均順位は 1.577 となり, 辞書ベース先頭削除で行う見出し生成は有効であることが示された. 見出し変化率を向上させるためには人手でより網羅性の高い辞書を作成する必要があるが, 人手で網羅するにはコストがかかり非現実的である.

ランク学習による見出し生成は, 勝率および平均順位の両方において比較手法の中で最も良い結果となった. 先頭切り出しよりも平均順位が良い結果となった辞書ベース先頭削除とランク学習の 2 つの手法の見出し変化率を比較したところ, ランク学習の方が見出し変化率が高くなっており, 高いカバレッジで適切な見出しを提供できることが示された.

提案手法であるランク学習の勝率と平均順位については, その他の比較手法それぞれに対して Wilcoxon 検定を実施し統計的に有意 ($p < 0.01$) であることを確認した.

6. まとめ

本研究では, クラウドソーシングを利用して Yahoo!知恵袋の見出しの教師データを作成し, それを用いたランク学習による見出し生成に取り組んだ. ランク学習による見出し生成手法は, 他の見出し生成手法と比較して質の高い見出しが生成できることがわかった. また, 人手で作成した辞書を利用した生成手法と比較して見出し変化率が高くなることが示され, サービスに反映する際のカバレッジをより高められることを確認した.

今後は, ランク学習の他にも深層学習を取り入れた手法の検討を行なっていく予定である. また, ランク学習によって生成した見出しのユーザーの回答率やクリック率への影響をオンラインで検証していくことも検討している. 本研究においてクラウドソーシングを利用して作成した投票のラベルは公開する予定である.

参考文献

[Bhaskar 13] Bhaskar, P.: Answering Questions from Multiple Documents – the Role of Multi-Document Summarization, in *RANLP 2013*, pp. 14–21 (2013)

[Heilman 10] Heilman, M. and Smith, N. A.: Good Question! Statistical Ranking for Question Generation, in *NAACL 21010*, pp. 609–617 (2010)

[Ishigaki 17] Ishigaki, T., Takamura, H., and Okumura, M.: Summarizing Lengthy Questions, in *IJCNLP 2017*, pp. 792–800 (2017)

[野口 15] 野口 正樹, 谷塚 太一, 小林 隼人：分散表現を用いたヤフー知恵袋の要約, 言語処理学会第 21 回年次大会 (2015)