分散表現を用いたヤフー知恵袋の要約

野口 正樹　　谷塚 太一　　小林 隼人
ヤフー株式会社
{manoguch, tyatsuka, hakobaya}@yahoo-corp.jp

1 はじめに

スマートフォンをはじめとするモバイル端末の所有率が増えるにつれ，スマートフォンを前提としたインターネットサービスを展開する必要が出てきた．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Q&A サイトや掲示板，SNS の投稿をまとめた “まとめサイト” などユーザジェネレートコンテンツ (User Generated Contents: UGC) を利用したサイトもスマートフォンから利用されるようになってきている．スマートフォンの場合には転送速度や表示領域の制約があるため，要点をまとめた短い文章でコンテンツを表す要約技術を用いたサービスの展開などを考える必要がある．

しかし，これまでの要約技術に関する評価において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実験は多く行われているが，UGC のように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．そこで，本論文では UGC サイトであるヤフー知恵袋 [5] に対して要約技術を適用しその効果を確認する．

本論文では重要文抽出による要約に取り組み，既存の手法に加え，単語の分散表現を使った手法を提案する．単語の分散表現は意味的な近さや関係性を表現できることで最近注目を浴びており [4]，単語だけでなくフレーズや文を固定長のベクトルで表現する取り組みが行われている [2]．重要文抽出において分散表現を利用した研究は，Kageback ら [1] の研究があるが，英語の評価用データを対象として実験を行っている．本論文では，日本語の UGC を対象としている点で Kageback らとは異なっている．

また，評価にはクラウドソーシングを用いた．クラウドソーシングでは一般のユーザによる定性評価をアンケート形式で手軽に収集することができる．クラウドソーシングを利用するメリットとして，専門家による評価ではなく一般のユーザの評価を得ることができることが挙げられる．

本論文の構成は以下の通りである．2 章で重要文抽出の既存の手法, 3 章で提案手法を述べる．4 章で実験設定および評価結果を述べ，5 章で考察を述べる．最後に 6 章で本論文をまとめる．

2 重要文抽出

文書 D に対する重要文の抽出結果をサマリ S，文書 D を構成する各文を $s_1, \ldots, s_n$ とすると，サマリ S は D の部分集合，すなわち $S \subset D = \{s_1, \ldots, s_n\}$ として表せる．元の文書 D に対するサマリ S の良さを $\mathrm{Score}_D(S)$，サマリのサイズを $\mathrm{Size}(S)$ とすると，重要文抽出問題は $\mathrm{Score}_D(S)$ の最大化問題として次のように定式化できる．

$$\begin{aligned} \max_{S \subset D} \quad & \mathrm{Score}_D(S) \\ s.t. \quad & \mathrm{Size}(S) \leq \ell \end{aligned} \tag{1}$$

ℓ はサマリサイズの上限値として与えるパラメータで，$\mathrm{Size}(S)$ には文字数，文の数などを返す関数を用いる．

2.1 TFIDF によるスコア

文の重要度を表す指標として TFIDF を利用する方法を説明する．各文 s がスコア $\mathrm{Score}(s)$ を持つとし，$\mathrm{Score}_D(S)$ を各文のスコアの線形和で定義する．

$$\mathrm{Score}_D(S) \quad = \quad \sum_{s \in S} \mathrm{Score}(s)$$

既存の手法として，各文に含まれる単語の TFIDF を利用して各文のスコアを計算するものがある [7]．ここでは，文 s に含まれる単語集合 W の各単語の TFIDF 値と文 s に対する重み $\mathrm{weight}(s)$ からスコア $\mathrm{Score}(s)$ を計算する．

$$\mathrm{Score}(s) \quad = \quad \mathrm{weight}(s) \sum_{w \in W} \mathrm{tfidf}(w).$$

本稿では weight(s) として，助詞と助動詞の割合を考慮したペナルティおよびその寄与度 C を用いる．

$$\text{weight}(s) \quad = \quad C\left(1 - \frac{\text{文内の助詞および助動詞数}}{\text{文全体の単語数}}\right)$$

このペナルティは助詞および助動詞が過剰に多く存在する文は価値が低いと仮定して設定した．このスコア関数は 0-1 ナップサック問題となるので，動的計画法で効率的に解くことができる．

3 提案手法

単語 w に対する分散表現 vec(w) を用いた文および文書の分散表現を使い，スコアに用いる手法を提案する．文 s に含まれる単語集合 W の各単語の分散表現 vec(w) を用い，文 s に対する分散表現および文集合 (文書)D に対する分散表現を次のように定義する．

$$\text{vec}(s) \quad = \quad \sum_{w \in W} \text{vec}(w). \tag{2}$$

$$\begin{aligned} \text{vec}(D) \quad &= \quad \sum_{s \in D} \text{vec}(s) \\ &= \quad \sum_{s \in D} \sum_{w \in W} \text{vec}(w). \end{aligned} \tag{3}$$

ここでは文の分散表現を用いる手法と文書の分散表現を用いる手法の 2 つを提案する．

3.1 文の分散表現によるスコア

文書 D に含まれる文 s に対する重要度スコア Score(s) に文書 D と s の分散表現のコサイン類似度を用いる．

$$\text{Score}_D(S) \quad = \quad \sum_{s \in S} \text{Cos}(\text{vec}(D), \text{vec}(s)). \tag{4}$$

ここで，2 つのベクトル V_1, V_2 を用いて Cos を次のように定義する．

$$\text{Cos}(V_1, V_2) \quad = \quad \frac{V_1 \cdot V_2}{|V_1|\ |V_2|}.$$

式 (4) を用いる場合，2.1 同様に 0-1 ナップサック問題として定式化できる．

3.2 文書の分散表現によるスコア

3.1 と同様に分散表現を利用するが，サマリ S に対する重要度スコア $\text{Score}_D(S)$ に文書 D とサマリ S の分散表現のコサイン類似度を用いる．

$$\text{Score}_D(S) \quad = \quad \text{Cos}(\text{vec}(D), \text{vec}(S)). \tag{5}$$

この場合，目的関数が線形和でないためナップサック問題として定式化できない．そのため本稿では貪欲法を用いる．貪欲法では，次のように 1 文ずつ文を選択していく．t 文選択した時点におけるサマリ集合を S_t とし，t 番目に選んだ文を s_t とすると, $S_t = S_{t-1} \cup \{s_t\}$ と表せる．ここで，$S_0 = \phi$ である．選択する文 s_t は，サマリ集合に加えた場合に最も $\text{Score}_D(S_t)$ が大きくなる文とする．すなわち，

$$s_t \quad = \quad \underset{s \in D \setminus S_{t-1}}{\text{argmax}}\ \text{Score}_D(S_{t-1} \cup \{s\}).$$

4 評価実験

提案手法の有用性を確認するため，Yahoo!クラウドソーシング [6] を用いて定性評価を行った．作業者は設問ごとに文章を読み、最も要点がまとまっているものを選択肢から選んでもらうタスクとした．

質問文に文集合 D を得るために，前処理として文分割を行った．句読点に加え，“！” や “？” 等の文末として用いられる記号を文境界の目印に用いた．表 2, 表 4 に文分割の例を示す．

また，本実験において制約条件として与える Size(S) には文の数を用い，3 文以内に収まるよう $l = 3$ とした．

4.1 実験設定

単語の分散表現を得るためにヤフー知恵袋の全質問文をトレーニングデータとして利用した．分散表現を得るためのツールとして word2vec[1] [3] を用い，モデルには CBoW，分散表現には 1,000 次元のベクトル，ウィンドウ幅を 8，ネガティブサンプル 25，12 スレッドで学習を行った．結果として 2,466,022 単語分の分散表現を得た．

TFIDF 値の計算には，タスクの簡単化のため 200 文字〜400 文字からなるヤフー知恵袋の質問文をランダムに抽出した 2,000 文書を用いた．

以降，本章で作成した TFIDF，分散表現を用いて 2.1 章の手法で抽出した手法を *TFIDF*，3.1 章の手法で抽出した手法を *SenVec*，3.2 章の手法で抽出した手法を *DocVec* と呼ぶ．各手法で抽出した重要文の例を表 3，表 5 に示す．

4.2 評価

前述の手法にて抽出した重要文を選択肢として提示し，評価の際には元文書となる質問文とともに，各手

[1] バージョン 0.1c を用いた．

法によって生成された要約文を選択肢として提示し，最も要点がまとまっているものを最も良い抽出文としてユーザに選択してもらった．

システム上の文字数制限やタスクの選択肢が長すぎる場合の回答率の低下に配慮し，今回掲載したタスクは 460 文書で，各文書ごとに 5 ユーザに評価してもらい，合計で 2,300 回答を得た．約 2 時間でタスクが完了し，費用は 4,600 円であった．

評価結果を表 1 に示す．異なる手法で同一の重要文が生成される場合があるため，その場合にはどちらの手法も選択されたものとみなして評価を行った．

表 1: クラウドソーシングによる評価結果

手法	選ばれた数	割合
TFIDF	655	28.5%
SenVec	1,106	48.1%
DocVec	671	29.2%

5 考察

表 1 から，本実験での条件設定では知恵袋の質問文に対する要約において，単語の分散表現を用いて文の重要度を決める手法がより良い要約を作れることが分かった．特に TFIDF 値を文の重要度に用いた *TFIDF* と分散表現を使った元文書とのコサイン類似度を文の重要度に用いた *SenVec* とを比較した場合，*SenVec* がより良い結果となった．これは分散表現を用いることで意味的に元文書に近い結果を得ることができていたからだと考えられる．

分散表現を使って抽出文と元文書とのコサイン類似度を重要度として用いた *DocVec* は *TFIDF* とあまり変わらず，同様に分散表現を用いた *SenVec* に及ばないという結果となった．*SenVec* では元文書の分散表現に近い文を優先して選択するため，*SenVec* の要約は意味的に近い文が含まれる．一方，*DocVec* では元文書の分散表現に近くなるように文を選択するため，*DocVec* では要約に含まれる文間の意味的な近さは考慮されない．このため，本実験で用いた $\ell = 3$ という条件のもとで *DocVec* の要約は意味的に遠い 1 文が含まれることがあり，これがまとまりがあまりないと判断された可能性がある．したがって，例えば，$\ell = 5$ などの条件に変更した際には意味的にまとまった 3 文と 2 文が選択されるなど，まとまり具合が変わり評価が変わることが予想される．

6 まとめ

本論文では単語の分散表現を用いる重要文抽出方法を提案した．ヤフー知恵袋の質問文に対して適用させた結果クラウドソーシングで評価し，従来の TFIDF を用いた抽出文に比べ良好な抽出結果となっていることを示した．

今回の評価に用いた元文書の文字数は 200-400 と比較的短いものを利用した．そのため，サービスへ適用するにあって，より長い質問文の場合の検証やサービスに適した制約条件を見つけることなどが課題として挙げられる．また，手法そのものの良さを検証するため，元文書に対してどの文が重要であるのか正解データを作り評価する必要がある．

本論文で提案した分散表現を用いる手法は適用先のドメインを限定するものではない．そのため，ヤフー知恵袋で学習した分散表現を別の UGC にも適用が可能である．今後は，様々なドメインのデータで学習し，提案手法の効果を検証していきたい．

参考文献

[1] Mikael Kageback and Devdatt Dubhashi Olof Mogren, Nina Tahmasebi. Extractive Summarization using Continuous Vector Space Models. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC 2014)*, pp. 31–39. Association for Computational Linguistics, 2014.

[2] Quoc Le and Tomas Mikolov. Distributed Representations of Sentences and Documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML 2014)*, pp. 1188–1196. JMLR, 2014.

[3] Thomas Mikolov. word2vec: Tool for computing continuous distributed representations of words, 2013. `https://code.google.com/p/word2vec/`.

[4] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26*, pp. 3111–3119. Curran Associates, Inc., 2013.

[5] Yahoo!知恵袋. `http://chiebukuro.yahoo.co.jp/`.

[6] Yahoo!クラウドソーシング. `http://crowdsourcing.yahoo.co.jp/`.

[7] 平尾努, 鈴木潤, 磯崎秀樹. 最適化問題としての文書要約. 人工知能学会論文誌 Vol.24 (2009) No.2, pp. 223–231. 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所, 2009.

表 2: 元文書と文分割の例 1

元文書
新築します。 現在間取り打ち合わせ中です。 5.4 帖の和室があり、間取り的に 1500 巾、奥行き 1000 巾の押入れ 1 つしか作れません。 お雛様を飾りたいと要望したら、ハウスメーカーからは 吊り押入れを提案されました。 押入れには、布団二組とお雛様を収納したいと思っています。 スペースがあれば座布団 4 枚も収納したいです。 吊り押入れ 1 つだけあって、その下にお雛ま様を飾るのは どうなんでしょうか？ 吊り押入れは重さの制限はあるのでしょうか？ 吊り押入れをやめて、普通の押入れにして お雛様を飾りたいときは畳の上に置く方が良いと思いますか？
文分割
新築します。 現在間取り打ち合わせ中です。 5.4 帖の和室があり、間取り的に 1500 巾、奥行き 1000 巾の押入れ 1 つしか作れません。 お雛様を飾りたいと要望したら、ハウスメーカーからは吊り押入れを提案されました。 押入れには、布団二組とお雛様を収納したいと思っています。 スペースがあれば座布団 4 枚も収納したいです。 吊り押入れ 1 つだけあって、その下にお雛ま様を飾るのはどうなんでしょうか？ 吊り押入れは重さの制限はあるのでしょうか？ 吊り押入れをやめて、普通の押入れにしてお雛様を飾りたいときは畳の上に置く方が良いと思いますか？

表 3: 手法ごとの抽出結果の例 1

TFIDF
現在間取り打ち合わせ中です。 5.4 帖の和室があり、間取り的に 1500 巾、奥行き 1000 巾の押入れ 1 つしか作れません。 お雛様を飾りたいと要望したら、ハウスメーカーからは吊り押入れを提案されました。
SenVec
お雛様を飾りたいと要望したら、ハウスメーカーからは吊り押入れを提案されました。 吊り押入れ 1 つだけあって、その下にお雛ま様を飾るのはどうなんでしょうか？ 吊り押入れをやめて、普通の押入れにしてお雛様を飾りたいときは畳の上に置く方が良いと思いますか？
DocVec
吊り押入れをやめて、普通の押入れにしてお雛様を飾りたいときは畳の上に置く方が良いと思いますか？ スペースがあれば座布団 4 枚も収納したいです。 お雛様を飾りたいと要望したら、ハウスメーカーからは吊り押入れを提案されました。

表 4: 元文書と文分割の例 2

元文書
近く、誕生日がありケーキを手作りしようと思っています。せっかくの誕生日なのでホールで作ろ>うと思っています。（直径が 12 センチ程の小さいやつです） 土台のスポンジ自体は購入してデコレーションだけをしようと考えています。彼に渡す当日を含め前後の日が全て仕事の為、スポンジを焼く時間が無いためそこは既製品のスポンジに頼ろうと思ってます。 今までケーキ作りはおろか、お菓子作りもしたことがないのでケーキのデコレーションをするためにどんな道具・材料から揃えたらいいか分かりません。また飾り付けのアイデアも浮かばず、どこから手を付けていいのか分からないのでケーキの飾り付けの良い見本になるようなサイトなどがあれば教えてください。 できれば素人でも簡単にできる。などであれば大変助かります。
文分割
近く、誕生日がありケーキを手作りしようと思っています。せっかくの誕生日なのでホールで作ろうと思っています。 （直径が 12 センチ程の小さいやつです） 土台のスポンジ自体は購入してデコレーションだけをしようと考えています。 彼に渡す当日を含め前後の日が全て仕事の為、スポンジを焼く時間が無いためそこは既製品のスポンジに頼ろうと思ってます。 今までケーキ作りはおろか、お菓子作りもしたことがないのでケーキのデコレーションをするためにどんな道具・材料から揃えたらいいか分かりません。 また飾り付けのアイデアも浮かばず、どこから手を付けていいのか分からないのでケーキの飾り付けの良い見本になるようなサイトなどがあれば教えてください。 できれば素人でも簡単にできる。 などであれば大変助かります。

表 5: 手法ごとの抽出結果の例 2

TFIDF
土台のスポンジ自体は購入してデコレーションだけをしようと考えています。 彼に渡す当日を含め前後の日が全て仕事の為、スポンジを焼く時間が無いためそこは既製品のスポンジに頼ろうと思ってます。 今までケーキ作りはおろか、お菓子作りもしたことがないのでケーキのデコレーションをするためにどんな道具・材料から揃えたらいいか分かりません。
SenVec
彼に渡す当日を含め前後の日が全て仕事の為、スポンジを焼く時間が無いためそこは既製品のスポンジに頼ろうと思ってます。 今までケーキ作りはおろか、お菓子作りもしたことがないのでケーキのデコレーションをするためにどんな道具・材料から揃え>たらいいか分かりません。 また飾り付けのアイデアも浮かばず、どこから手を付けていいのか分からないのでケーキの飾り付けの良い見本になるようなサイトなどがあれば教えてください。
DocVec
今までケーキ作りはおろか、お菓子作りもしたことがないのでケーキのデコレーションをするためにどんな道具・材料から揃えたらいいか分かりません。 また飾り付けのアイデアも浮かばず、どこから手を付けていいのか分からないのでケーキの飾り付けの良い見本になるようなサイトなどがあれば教えてください。 せっかくの誕生日なのでホールで作ろうと思っています。