ARG WI2 No.13, 2018
WI2-2018-21

単語の分散表現及び質問と回答における単語の共起頻度を利用したFAQ検索手法

奥野 翔太　　荒木 健治

北海道大学大学院情報科学研究科

s-okuno2@eis.hokudai.ac.jp　　*araki@ist.hokudai.ac.jp*

概要　本稿では，単語の分散表現及びFAQの質問文と回答文に含まれる単語の共起頻度を利用したFAQ検索手法を提案する．本手法の特徴は，ユーザの入力に含まれる内容語と単語の分散表現を用いて獲得した内容語に類似する単語の利用，及びFAQの質問文と回答文に出現する単語の共起頻度に基づき，クエリに対して回答文で頻出する単語の利用により，様々な自然言語での質問に対してユーザの所望するFAQを高精度に検索できることである．さらに，FAQの質問スコアと回答スコアの重要性の違いを考慮し，回答スコアに重みを加えた手法についても提案し考察を行う．本手法に基づく実験システムを作成し，評価実験を行なった結果，我々の従来手法と比較してMRRの値が最大6.8ポイント改善され，提案手法の有効性が確認された．

キーワード　情報検索，FAQ，FastText

1 はじめに

大企業や官公庁などのWebサイトには，これまでにユーザから多く寄せられた質問とその回答をまとめた“よくある質問”(FAQ)が掲載されていることが多く，疑問を抱えたユーザがFAQを見ることで，疑問を解消することができる．また，コールセンターなどではユーザの問い合わせに対して迅速に対応するために，想定される質問とその回答をまとめたものをFAQとして蓄積していることや，知識の浅い新人オペレーターの補助としてFAQを利用している[1]．掲載されているFAQの件数は企業によって様々であり，少量であればユーザは所望するFAQを容易に見つけることができるが，件数が膨大になるとユーザの所望するFAQの検索は困難である．そのため，ユーザの検索要求を満たすFAQを適切に検索できるシステムの需要が高まっており，様々な研究が行われている[2]．

また，近年では雑談だけでなくユーザの質問にも返答可能なAIチャットボットが登場し[3]，高度なFAQ検索や質問応答の技術が求められている．

そこで，我々はそのような技術に応用可能なFAQ検索手法を目指し，様々な自然言語の質問に対してユーザの所望するFAQが検索可能な手法の開発を試みた．本稿では，単語の分散表現から獲得した類似語やFAQの質問文と回答文に含まれる単語の共起頻度を利用したFAQ検索手法を提案する．

2 関連研究

既存のFAQ検索手法の問題点として，同義語や表記ゆれによる検索漏れの問題があり，検索語の表記のみを利用した全文検索手法などでは意味的に同じでもユーザの入力する質問文の表現によって検索結果が異なってしまう．この問題に対して，賈らの研究[4]ではEarth Mover’s Distanceや依存構文を用いて文に含まれる単語同士の関係性や文法規則を考慮した文間関連度計算手法を提案し，文書検索や辞書検索など自然言語を処理する検索システムへの応用を行なっている．

また，川田らの研究[5]では，項を持たない動詞の含意ペアに簡単なルールを適用することで，項が1つ以上のより複雑な言語パターンの含意ペアデータを生成し，言語パターンの数を拡張する手法を提案している．これによりFAQ検索や質問応答のシステムにおいて同じ回答を得られる質問の表現のパターン数を増やすことが期待できる．

一方で，牧野らの研究[6]では，コールセンターへの問い合わせ履歴を利用して自動生成した学習データを用いて，問い合わせを対応するFAQへ分類する文書分類器を学習し，その文書分類器の出力をランキング学習の素性として用いるFAQ検索手法を提案している．

これらの研究と比較して，本研究では言い換え表現や表記ゆれに対して単語の分散表現から獲得した類似語を用いる点や，FAQのスコア算出方法においてFAQの質問文と回答文における単語の出現頻度の違いを考慮している点で異なる．

表1 学習コーパスの詳細

コーパス名	内容	データ量
S	ソフトバンクの FAQ データ	11.9MB
SDA	ソフトバンク, ドコモ, au の FAQ データ	20.7MB
SDAR	ソフトバンク, ドコモ, au, 楽天モバイルの FAQ データ	27.2MB
W	Wikipedia のダンプデータ	6.6GB

3 単語の分散表現に関する予備実験

これまでに我々が行なった研究[7]として単語の分散表現を用いた FAQ 検索手法がある. 単語の分散表現を利用してユーザの入力に含まれる内容語に類似する単語を獲得し, 検索クエリとして利用することで, 様々な自然言語での質問に対して適切な FAQ を検索することができる. しかし, 従来研究では FAQ 検索に対してより有効な単語の分散表現の生成方法に関する考察は行われていない.

そこで, 本稿では新たに分散表現の学習コーパスを 4 種類用意し, それぞれの学習コーパスから 2 種類の分散表現生成ツールを用いて 8 種類の分散表現を生成した. 本章では, これらの分散表現を利用して我々の従来手法を用いて予備実験を行い, FAQ 検索に適した単語の分散表現についての考察を行う.

3.1 学習データ

単語の分散表現を生成するための学習コーパスとして, 従来研究ではソフトバンク[8]から収集した FAQ データを使用した. 本稿ではこの FAQ データに加えて, ドコモ[9], au[10], 楽天モバイル[11]の Web サイトからそれぞれ FAQ データを収集した. データ量はそれぞれ 11.9MB, 3.0MB, 5.8MB, 6.5MB であった. これらの FAQ データを組み合わせ, 2 種類の学習コーパスを作成した. また, FAQ データ以外の学習コーパスとの比較を行うため, Wikipedia のダンプデータ(6.6GB)[12]も学習コーパスとして利用した. これら 4 種類の学習コーパスの詳細を表1に示す. 以下, 本稿では各学習コーパスを表中のコーパス名で呼ぶ.

3.2 分散表現生成ツール

本稿では, 分散表現の生成ツールとして Word2Vec[13]と FastText[14]を使用した. これらのツールは学習コーパスに含まれる単語に言語概念ベクトルを付与することができる. そのため, 単語同士のベクトル間のコサイン類似度を計算することで, 単語同士の意味的な近さを数値化することができる. これに加えて, FastText では単語だけでなく subword と呼ばれる単語の文字 n-gram にもベクトルを付与することができるため, 動詞の活用形を考慮した単語ベクトル生成や, 表層的に似ている単語同士や未知語, 略語への対応に優れている. また, Word2Vec と比較して FastText は学習コーパスのデータ量が少なくても正確な分散表現を生成することができる. そのため, FAQ データのようなデータ量が比較的少ない学習コーパスにおいても正確な分散表現の生成が期待できる.

3.3 手法概要

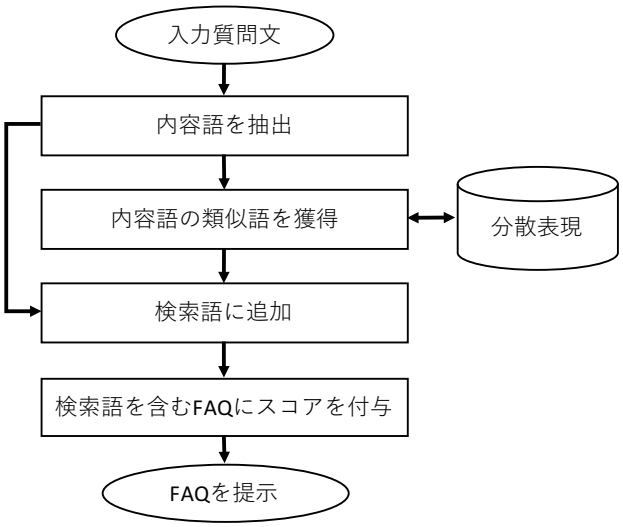

図 1 予備実験の手法の処理過程

予備実験の手法の処理過程を図1に示す. 以下, ユーザが入力する質問文を入力質問文と呼ぶ. 入力質問文に対して MeCab[15]で形態素解析を行い, 内容語を抽出して検索語$\{q_1, q_2, \cdots, q_m\}$として獲得する. ここで内容語のみを検索語としたのは, 入力質問文と FAQ の質問文の意味的な近さを計算する際に, 助詞や助動詞などの機能語の表層的な一致率が高いために正解でない FAQ が誤って上位に順位付けられてしまう場合が考えられるためである. 続いて, 単語の分散表現を用いて検索語の単語ベクトルとのコサイン類似度が閾値以上の単語を類似語$\{q'_1, q'_2, \cdots, q'_n\}$として獲得し, 検索語に追加する. これにより, 既存の FAQ 検索手法の問題点である同義語や表記揺れによる検索漏れの問題が解消できる.

検索語q_iが質問文Qに含まれていれば, 検索語q_iの質問文Qにおける tf-idf 値を検索語q_iのスコアとし, 質問

文Qのスコアに加算する. 全ての検索語で同様の作業を行い, 各検索語のスコアの総和を質問文Qのスコアとする. このとき, 入力質問文から抽出した内容語と比較して, 単語の分散表現により獲得した類似語は検索語としての重要性が低くなると考えられる. よって, 類似語q'_kのスコアはq'_kの tf-idf 値と, 検索語q_iのベクトルvec_{q_i}と類似語のベクトル$vec_{q'_k}$のコサイン類似度の積とする. 質問文Qのスコアを式(1)に示す.

$$Score_Q = \sum_i^m tfidf_{q_i,Q} + \sum_i^m \sum_k^n tfidf_{q'_k,Q} \cdot cos\left(vec_{q_i} \cdot vec_{q'_k}\right) \quad (1)$$

また, 回答文Aに対しても質問文Qと同様の手順でスコア付与を行う. 回答文Aのスコアを式(2)に示す.

$$Score_A = \sum_i^m tfidf_{q_i,A} + \sum_i^m \sum_k^n tfidf_{q'_k,A} \cdot cos\left(vec_{q_i} \cdot vec_{q'_k}\right) \quad (2)$$

式(3)に示すように質問文Qのスコア$Score_Q$と回答文Aのスコア$Score_A$の和を FAQ のスコアとする.

$$Score = Score_Q + Score_A \quad (3)$$

全ての FAQ について同様の作業を行い, スコアが上位 10 件の FAQ をユーザに提示する.

3.4 実験設定

実験データにはソフトバンクの Web サイトから収集した 8,013 件の FAQ を使用した. 単語の分散表現から類似語を獲得して検索語に追加する基準となるコサイン類似度の閾値は, 0.50~0.75 で 0.05 ずつ変更して実験を行なった.

分散表現の学習方法は Word2Vec, FastText 共に skip-gram, 単語ベクトルの次元数は Word2Vec で 200, FastText で 100 に設定して実験を行なった.

被験者は 20 代の理系大学院生 9 名(男性 8 名, 女性 1 名)である. 被験者にはソフトバンクの FAQ カテゴリを参考に, 10 件の質問文を自由に入力してもらい, それぞれの質問に対してシステムが提示した上位 10 件の FAQ が正解であるかどうかの判定を行なった. 計 90 件の入力質問文に対する評価を行なった.

評価指標には MRR（Mean Reciprocal Rank）及び Precision@N（Pre@N）を用いた. MRR はシステムに提示された上位 5 件の FAQ のうち最も上位にある正解の FAQ の順位の逆数の平均値である. P@N は正解の FAQ が N 位以上にある割合である.

3.5 実験結果

単語の分散表現の生成ツールとして Word2Vec を使用した際の実験結果を表 2 に, FastText を使用した際の実験結果を表 3 に示す. 表中の値は, 各組み合わせにおいて MRR が最大となる閾値で実験した際の結果である. それぞれの組み合わせにおいて MRR が最大となった時の閾値は, Word2Vec を使用して学習コーパスが SDA の場合は 0.50, SDAR の場合は 0.55, FastText を使用して学習コーパスが SDAR の場合は 0.70, それ以外の場合は 0.60 となった.

表 2 Word2Vec の場合

コーパス	MRR	Pre@1	Pre@5	Pre@10
S	0.443	0.333	0.644	0.711
SDA	0.450	0.367	0.633	0.700
SDAR	0.451	0.344	0.633	0.733
W	0.420	0.300	0.611	0.700

表 3 FastText の場合

コーパス	MRR	Pre@1	Pre@5	Pre@10
S	0.447	0.344	0.622	0.744
SDA	0.455	0.333	0.678	0.733
SDAR	0.458	0.356	0.644	0.722
W	0.443	0.322	0.656	0.733

表 2, 3 に示すように, Word2Vec, FastText どちらの場合においても学習コーパスが SDAR の場合に MRR の値が最も高くなり, W の場合に MRR の値がもっとも低くなった. また, 全ての学習コーパスにおいて Word2Vec よりも FastText で分散表現を生成した場合の方が MRR の値が高くなった.

3.6 考察

FAQ データを学習コーパスとした場合, データ量が多いほど MRR の値が高くなり, システムの精度が向上している. この理由として, ソフトバンクの FAQ データに他社の FAQ データを加えることで, 学習コーパスにおいて 4 社の FAQ に共通して出現する単語の出現頻度が増加し,「スマートフォン」や「通信量」といったモバイル FAQ に関連する単語においてより正確な分散表現が生成されたと考えられる.

また, 全ての学習コーパスにおいて Word2Vec と比較して FastText の方が MRR の値は高くなっている. これは FastText の単語だけでなく単語の文字 n-gram にもベクトルを付与するという特徴により,「アプリ」と「アプリケーション」といった表層的に類似している略語などのベクトル化の精度が向上していることが理由として挙げられる. この結果から, Word2Vec よりも FastText の方が FAQ 検索に使用する分散表現生成ツールとして優れていることがわかる.

続いて Wikipedia のダンプデータを学習コーパスとした場合に精度が低くなった理由について考察する. 予備実験の手法では分散表現から獲得した類似語が多く

なりすぎて入力質問文から抽出した内容語の検索語としての重要度が低くなるのを防ぐため，各内容語につき獲得できる類似語の数の上限を 10 個に限定している．そのため，入力質問文中の内容語とのコサイン類似度が上位 10 位以内の単語の中にモバイル FAQ に関連する単語が出現しない場合，検索語に類似語を追加しても FAQ 検索には有効ではない．実際に学習コーパスを Wikipedia のダンプデータとした場合，一般的には意味が似ており，関連性の高い単語ではあるがモバイル FAQ に関連のない単語が類似語として獲得されることが多かった．このことから，Wikipedia のダンプデータを学習コーパスとして生成した分散表現は正確なベクトル化が行われていたとしても FAQ 検索には不向きであると考える．

4 提案手法

これまでの我々の手法では，入力質問文に含まれる内容語とその類似語は FAQ の質問文だけでなく回答文にも出現しやすいと仮定し，質問文と回答文で同じ方法でスコアを算出していた．しかし，それらの単語が FAQ の質問文に出現しやすいことは明らかであるが，回答文にも出現しやすいとは限らない．

また，我々の従来研究において，質問文と回答文における単語の出現頻度の違いについての考察は行われていない．実際，FAQ における出現回数が 100 回以上の単語のうち，質問文と回答文での出現割合の比率が 5 倍以上である単語は 225 個存在し，これは出現回数が 100 回以上の単語全体の 22.5%を占めている．また，「どの」や「何」のような疑問詞や「教えて」のような質問する際によく用いられる単語は質問文における出現回数が多い．逆に，「お願い」や「ご覧」のような丁寧な表現や，「手順」や「こちら」のような解決法を示すような単語は回答文における出現回数が多くなっている．

そこで，本稿では質問文と回答文における単語の出現頻度の違いに着目した FAQ の回答文のスコアの算出方法を新たに提案する．本手法の処理過程を図 2 に示す．

4.1 質問文スコアの算出方法

質問スコアの算出方法は我々の従来手法における質問スコアの算出方法と同じである．これは，**3.3** で述べた．

4.2 共起語リスト作成

回答文スコア算出の事前準備として，共起語リストを生成する．ここで，共起語とは質問文に出現する単語に対して回答文における共起回数が最多となる単語のことをいう．

各 FAQ において質問文に出現する単語と回答文に出現する単語それぞれのペアの共起回数を算出し，三

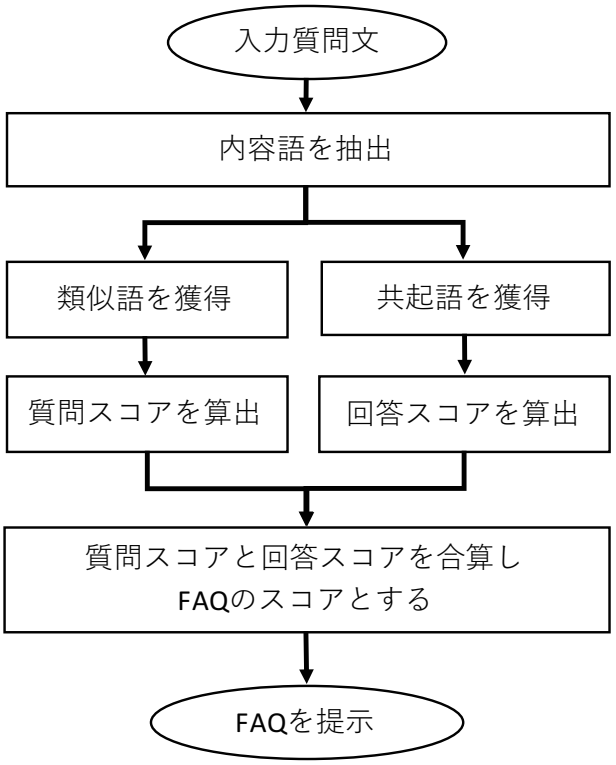

図 2 本手法の処理過程

つ組{質問文の単語,回答文の単語,共起回数} を生成する．このとき，助詞や助動詞などの機能語は出現回数が極端に多いためストップワードとした．また，賈らの研究[2]を参考に，内容語の中でも文の意味を表すことのできない出現回数の多い単語もストップワードとした．具体的には「する」，「下さる」，「ため」，「ある」といった単語である．全ての FAQ において同様の三つ組を作成し，質問文と回答文の単語ペアごとに各 FAQ における共起回数を合算する．このことにより，質問文に出現する単語と回答文に出現する単語の共起性が可視化できる．また，ある単語が質問文に出現していれば，回答文にはどのような単語が出現しやすいかということがわかる．

4.3 回答スコアの算出方法

入力質問文に含まれる内容語$\{q_1, q_2, \cdots, q_m\}$に対して，共起語リストを用いてそれぞれの内容語の共起語$\{c_1, c_2, \cdots, c_m\}$を獲得する．共起語$c_i$が回答文Aに含まれている場合，共起語$c_i$の回答文Aにおける tf-idf 値を共起語$c_i$のスコアとし，回答文Aのスコアに加算する．全ての共起語において同様の作業を行い，各共起語のスコアの総和を回答文Aのスコアとする．回答文Aのスコアを$Score'_A$とするとき，計算式は式(4)のようになる．

$$Score'_A = \sum_i^m tfidf_{c_i,Q} \quad (4)$$

質問文Qのスコア$Score_Q$と回答文Aのスコア$Score'_A$を合わせたものを FAQ のスコア$Score'$とする．

$$Score' = Score_Q + Score'_A \quad (5)$$

5 評価実験

5.1 実験方法

実験データ, 分散表現の学習パラメータ, 実験方法, 評価指標は予備実験と同様の方法で行った. 分散表現生成ツールはFastText, 学習コーパスはSDA, SDARを使用して比較を行った. また, 分散表現を用いて獲得する類似語のコサイン類似度の閾値は, 学習コーパスが SDA の場合は 0.60, SDAR の場合は 0.70 とした.

5.2 実験結果

評価実験の結果を表 4 に示す. 表 3 の予備実験の結果と比較して MRR の値は学習コーパスが SDA の場合に 1.9 ポイント, SDAR の場合に 0.9 ポイント向上している. また, 学習コーパスが SDA の場合は表 3 の結果と比較して MRR, Pre@1 において有意水準 0.05 で有意差が確認された.

表 4 評価実験の結果

コーパス	MRR	Pre@1	Pre@5	Pre@10
SDA	0.474	0.367	0.656	0.711
SDAR	0.467	0.367	0.611	0.689

5.3 質問文スコアと回答文スコアの重要度

回答文スコアの算出方法を変更することによりシステムの精度は向上した. しかし, 提案手法においてユーザの入力質問文に含まれる内容語とその類似語を用いる質問文スコアと, 共起語を用いる回答文スコアでは, スコアとしての性質や大きさが異なるため, 単純に合算した値を FAQ のスコアとするのは不適切である. そこで, 回答文スコアに 0.0~1.5 で 0.1 ずつ変更して重みwを加え, 実験を行なった. FAQ スコアを$Score''$とするとき, 計算式は式(6)のようになる.

$$Score'' = Score_Q + Score'_A * w \quad (6)$$

実験結果を表 5 に示す. 表中の値は, それぞれの学習コーパスにおいて MRR の値が最大となる重みwで実験した際の結果である. MRR の値が最大になった時の重みwは, SDA の場合は 0.4, SDAR の場合は 0.7 となった. また, 図 3 は重みwを変化させた場合の各学習コーパスでの MRR の値の変化を示したグラフである.

表 5 重みwを変更した場合の実験結果

コーパス	MRR	Pre@1	Pre@5	Pre@10
SDA	0.511	0.422	0.656	0.711
SDAR	0.492	0.411	0.611	0.644

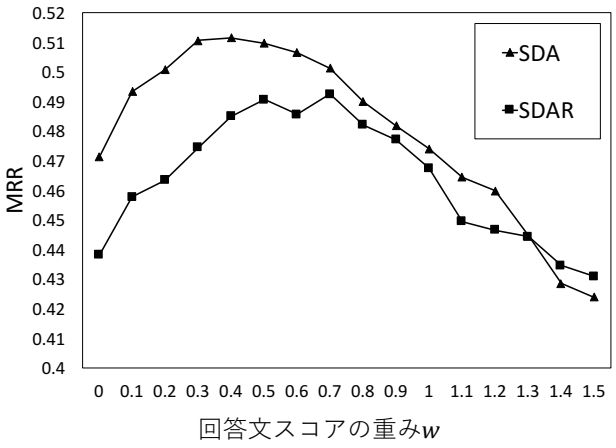

図 3 重みwを変更した場合の実験結果

表 4 の結果と比較して MRR の値は学習コーパスが SDA の場合に 3.7 ポイント, SDAR の場合に 2.5 ポイント向上している. また, どちらの学習コーパスにおいても表 4 の結果と比較して MRR, Pre@1 において有意準 0.05 で有意差が確認された.

6 考察

表 4 の結果を見ると, どちらの学習コーパスにおいても MRR と Pre@1 の値が予備実験の結果を上回っている. このことから, 入力質問文中の内容語とその類似語は回答文に出現しやすいとは言えず, FAQ の質問文と回答文に出現する単語の共起頻度を考慮した回答文スコアの算出方法の方が FAQ 検索には有効であり, ユーザの所望する FAQ を上位に順位付けできる.

学習コーパスが SDA の場合, 予備実験の結果と比較して MRR と Pre@1 の値は上がっているが, Pre@5 と Pre@10 の値が下がっている. この結果から, 5 位または 10 位以内に正解の FAQ が順位付けできている件数が減少してしまったと考えられる. しかし, 実際は評価指標に関わる順位内で正解の FAQ の順位が上昇した件数が 9 件あるのに対して順位が下降した件数が 7 件であり, 順位が上昇した件数の方が多い. また, 全体で見ても正解の FAQ の順位が上昇した件数が 26 件あるのに対して順位が下降した件数が 17 件であり, 順位が上昇した件数の方が多くなっている. 順位が上昇した件について調査したところ, 不正解の FAQ と比べて正解の FAQ の質問文スコアが小さくなってしまっているが, 回答文スコアが大きいために順位が上昇したケースが多く存在しており, 新しい回答文スコアの有効性が確認できた.

また, 予備実験の際には学習コーパスが SDAR の場合, SDA の場合よりも MRR の値が高くなっていたが, 今回の実験では SDA の場合の方が全ての指標におい

て上回っていた. このことから, 学習コーパスに他社の FAQ データを追加すればするほどより良い分散表現を生成することができ, システムの精度向上に繋がるとは一概に言えない.

そこで, 他社の FAQ データを学習コーパスに追加することで, ソフトバンクの FAQ 固有の単語にどのような影響があるのか調査を行なった. ソフトバンクの FAQ に出現する単語のうち,

- 品詞が固有名詞である
- 他社の FAQ に出現しない
- ソフトバンクの FAQ カテゴリに掲載されている

の 3 つの条件を全て満たす単語をソフトバンクの FAQ 固有の単語としたところ, 41 単語存在した. 学習コーパスが S の場合と SDA の場合で, ソフトバンクの FAQ 固有の単語同士のベクトル間のコサイン類似度を算出し, 分散表現間での単語ペアごとのコサイン類似度の差を算出したところ, 平均値は 0.0312, 中央値は 0.0260 となった. 同様の作業を学習コーパスが S の場合と SDAR の場合で行なった結果, 平均値が 0.0403, 中央値が 0.0372 となった. この結果から, 学習コーパスに他社の FAQ データを追加すればするほど, ソフトバンクの FAQ 固有の単語のベクトル化への影響が大きくなるため, 学習データに追加する他社の FAQ データはソフトバンクの FAQ データより少ない方が良いと考えられる.

表 5 の結果を見ると, どちらの学習データにおいても表 4 の結果と比較して MRR,Pre@1 の値が上昇している. このことから, 質問文スコアと回答文スコアでは FAQ スコアとしての重要度が異なり, 質問文スコアの方が重要であることがわかる. また, 回答文スコアに重みwを加えることで順位が上昇した場合についての調査を行なった. その結果, 不正解の FAQ と比べて正解の FAQ の質問文スコアは大きいが回答文スコアが小さくなってしまっている場合に, 重みwを加えて回答文スコアの大きさを小さくすることで正解の FAQ の順位が上がるというケースが多く存在しており, 回答文スコアに重みwを加えることの有効性が確認できた.

7 おわりに

本稿では, 単語の分散表現を用いて獲得した類似語及び質問文と回答文の単語の共起頻度を利用した FAQ 検索手法を提案した. また, 質問文スコアと回答文スコアの重要度の違いに着目し, 回答文スコアに重みを加えて実験を行い, 比較した. その結果, 我々の従来手法と比較して MRR の値が最大 6.8 ポイント改善され, 提案手法の有効性が確認された.

また, 予備実験と評価実験を通して FAQ 検索により有効な単語の分散表現の生成方法について考察を行なった. その結果, 学習コーパスに他社の FAQ データを追加することでより良い単語の分散表現が生成されるが, 他社の FAQ データが多すぎることによる単語の分散表現への悪影響も確認された.

今後は, いかなる手法においても 10 位以内に順位付けできていない質問文が 10%(90 件中 9 件)存在するため, それらの適切なスコアの算出方法について検討する必要がある. 具体的には, 質問カテゴリを考慮した質問文スコアの算出方法の検討や, 共起頻度を考慮した回答スコアの算出方法の更なる改良が挙げられる.

参考文献

[1] 長谷川友治, 大囿忠親, 伊藤孝行ほか:コールセンターにおける質問応答データの FAQ 作成支援システムの試作, 日本ソフトウ会第 20 回記念大会論文集, pp. 7-78, 2003.

[2] 木村英志, 高島俊哉, 重岡知昭ほか:文書カテゴリを利用した文書クラスタリングのコールセンター FAQ 改善への適用, 第 76 回全国大会講演論文集, pp. 485-486, 2014.

[3] 川端貴幸, 佐藤一誠:意味と表記の組み合わせによる用例ベースの質問応答モデル, 人工知能学会全国大会論文集 2017 年度人工知能学会全国大会(第 31 回)論文集, p. 1B2OS25b4, 2017.

[4] 賈珍磊, 吉村枝里子, 土屋誠司ほか:日本語における依存構文と EMD に基づいた文間の関連度計算手法, 研究報告知能システム(ICS), pp.1-8, 2016.

[5] 川田拓也, Kloetzer Julien, 鳥澤健太郎ほか:質問応答システムのための含意パターンペアの生成, 言語処理学会第 21 回年次大会発表論文集, pp. 159-162, 2015.

[6] 牧野拓哉, 野呂智哉, 岩倉友哉:自動生成した学習データを用いた文書分類器に基づく FAQ 検索システム, 自然言語処理 24.1, pp.117-134, 2017.

[7] 奥野翔太, 荒木健治:単語の分散表現により獲得した類似語を用いた FAQ 検索システムの性能評価, 第 10 回 Web インテリジェンスとインタラクション研究会資料, pp.23-24, 2017.

[8] ソフトバンク よくあるご質問(FAQ) https://www.softbank.jp/support/faq/

[9] NTTdocomo よくあるご質問(FAQ) http://faq.nttdocomo.co.jp/faq/p/top.do

[10] au よくあるご質問 https://www.au.com/support/faq/

[11] 楽天モバイル ヘルプページ https://mobile.faq.rakuten.ne.jp/app/home

[12] WikipediaDownloads https://dumps.wikimedia.org

[13] Mikolov, T., Sutskever, I., Chen, K., et al.: Distributed representations of words and phrases and their compositionality, In Advances in neural information processing systems, pp. 3111-3119, 2013.

[14] Bojanowski, P., Grave, E., Joulin, A., et al.: Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606, 2016.

[15] Kudo, T., Yamamoto, K., Matsumoto, Y.: Applying Conditional Random Field to Japanese Morphological Analysis, Proc. of the 2004 Conference on Emprical Method in Natural Language Processing, pp. 230-237, 2004.