ARG WI2 No.10, 2017
WI2-2017-6

単語の分散表現により獲得した類似語を用いた FAQ 検索システムの評価性能

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　近年，Webサイトに FAQを用意する企業が増加し，ユーザが FAQを利用することで自己解決を促している．しかし，自己解決できなかったユーザは多く存在し，FAQ を利用することのユーザの満足度は低い．したがって，ユーザの所望する FAQ を適切に検索するシステムの必要性が高まっている．本稿では，ユーザの入力に含まれる内容語に加え，単語の分散表現を用いて内容語に類似する表現を獲得し，それらの表現を検索語とした FAQ 検索手法を提案する．本手法に基づく実験システムを作成し，FAQ を対象として評価実験及び考察を行った結果について述べる．

キーワード　情報検索，Word2Vec，tf-idf

1 はじめに

Web 上には FAQ サイトやコミュニティ型 Q&A サイトが多く存在し，ユーザに問題が生じた際にそれらを利用することで自己解決を促している．しかし，FAQ サイトには大量の FAQ が掲載されているため，ユーザが所望する FAQ を検索できない，あるいは検索できたとしても多くの時間を要してしまうなどの問題がある．よって FAQ を利用することのユーザの満足度は低い．そこで，ユーザの所望する FAQ を適切に検索できるシステムの需要が高まっている．本稿では，Word2Vec[1]を用いて単語の分散表現を生成し，検索語に類似する表現を分散表現により獲得して検索語に追加することで，ユーザの質問に対して適切な FAQ を検索する手法を提案する．

2 関連研究

FAQ 検索システムの問題点として，同じ意味を持つが表記の異なる単語である言い換え表現や表記ゆれによる検索漏れの問題がある[2][3]．このような単語ペアが意味的に同じかどうかを判定できないと適切な FAQ を検索することは難しい．

この問題に対して，質問文と回答文を対訳コーパスとみなして作成した統計的翻訳モデルにより，ユーザの質問文と FAQ の質問文に含まれる単語ペアが同義かどうかを推定する研究[4]がある．また，動詞の含意ペアから複雑な言語パターン間の含意ペアデータを生成することで，同じ回答を得られる質問の数を増やす研究[5]がある．これらの研究と比較して，本研究は言い換え表現や表記揺れに対して単語の分散表現から獲得した類似表現を用いる点が異なる．

3 提案手法

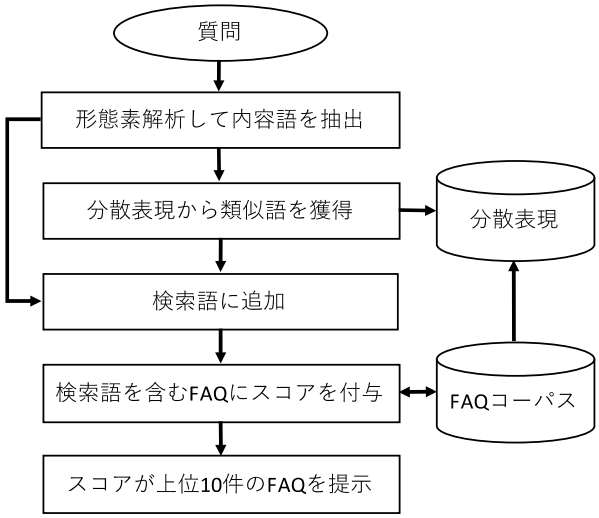

図 1　処理の流れ

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 を用いて形態素解析を行い，検索語$\{q_1, q_2, \cdots, q_m\}$を抽出する．FAQ の質問文と入力質問文を比較する際に，助詞や助動詞などの機能語の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．よって機能語をストップワードとし，内容語のみを検索語とする．

本研究では，検索語に加えて検索語に類似する表現を含む FAQ に，ユーザの所望する内容が記載されやすいと仮定する．よって，Word2Vec により作成した単語の分散表現を用いて検索語の分散表現とのコサイン類

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．

次に，FAQ コーパスにある FAQ をそれぞれ質問文と回答文に分類する．それぞれの文書で形態素解析を行い，文書中の各単語の tf-idf 値を算出する．

検索語q_iが質問文Qに含まれている場合，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のスコアに加える．すべての検索語において同様の作業を行い，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検索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の分散表現vec_iと類似語の分散表現vec_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．

$$Score_Q = \sum_i tfidf_{i,Q} + \sum_i \sum_k tfidf_{k,Q} \cdot cos(vec_i \cdot vec_k) \quad (1)$$

また，質問文Qに対応する回答文Aに検索語が多く含まれる場合が考えられる．そのような FAQ にもユーザの所望する内容を多く記載していると仮定し，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．

$$Score_A = \sum_i tfidf_{i,A} + \sum_i \sum_k tfidf_{k,A} \cdot cos(vec_i \cdot vec_k) \quad (2)$$

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

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

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

4 評価実験

実験データとして，ソフトバンクモバイルの FAQ サイト[6]から 8,013 件の FAQ を取得し，FAQ コーパスを作成した．FAQ コーパスには FAQ 固有の単語が多く存在するため，Word2Vec の学習データとして用いることで固有表現の類似語を適切に獲得できると考えられる．よって，Word2Vec の学習には FAQ コーパスを使用した．データ量は約 13MB で，語彙数は約 6,000 語であった．

類似語を分散表現により獲得する際のコサイン類似度の閾値は仮に 0.6 とした．入力質問文から抽出した内容語のみを検索語とした場合をベースラインとし，提案手法との比較を行った．

被験者は 20 代の理系大学生 3 名，理系大学院生 4 名の計 7 名である．被験者はソフトバンクモバイルの FAQ カテゴリを参考に，10 件の質問文を自由に入力した．それぞれ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った．計 70 件の入力質問文に対する評価を行った．

評価指標には MRR（Mean Reciprocal Rank）及び Precision@N（Pre@N）を用いた．MRR は上位 5 個の FAQ のうち最も上位にある正解の順位の逆数の平均値である．P@N は正解が N 位以上にある割合である．

結果を表 1 に示す．MRR, Pre@1, Pre@5 において提案手法がベースラインを上回る結果となった．

表 1 評価実験の結果

	MRR	Pre@1	Pre@5	Pre@10
提案手法	0.477	0.371	0.671	0.743
ベースライン	0.436	0.343	0.614	0.743

実験結果より，正解の FAQ には入力質問文中の内容語の類似語が含まれやすく，類似語のスコアが正解の FAQ のスコア加わることで正しい順位付けが行われたと推測できる．また，分散表現のコサイン類似度が 0.6 以上である単語ペアを確認すると，「どうして」と「なぜ」のような同じ意味を持つ単語ペアが多数存在し，言い換え表現や表記ゆれの問題を解消できていると確認された．しかし，低頻出語の分散表現が正しく学習できておらず，Word2Vec の学習データをソフトバンクモバイルの FAQ コーパスに限定するのは不十分であると言える．

5 おわりに

単語の分散表現を用いて獲得した単語を検索語に追加することで，ユーザの所望する FAQ を上位に順位付けられることを確認した．今後は，FAQ 検索システムに対してより有効な単語の分散表現の生成方法を検討する予定である．

参考文献

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

[2] 萩原正人，鈴木久美：意味的類似度を利用した日本語クエリ書き換え，言語処理学会第 15 回年次大会発表論文集，pp. 522-525, 2009.

[3] 山本和英：日本語の表記ゆれ問題に関する考察と対処，JAPIO YEAR BOOK 2015，pp. 202-205，2015.

[4] Xue, Xiaobing, Jiwoon Jeon, and W. Bruce Croft.: Retrieval Models for Question and Answer Archives, Proc. of ACM SIGIR Conference, pp. 475-482, 2008.

[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] ソフトバンクモバイル よくあるご質問(FAQ), http://faq.mb.softbank.jp/default.aspx.