

自動生成した質問に基づく質問応答学習手法の提案と評価

牧野拓哉^{1,2}, 野呂智哉¹, 吉川和^{1,2}, 岩倉友哉^{1,2}, 関根聡², 乾健太郎^{2,3}

株式会社富士通研究所¹

国立研究開発法人 理化学研究所 革新知能統合研究センター²

東北大学³

{makino.takuya,t.noro,y.hiyori,iwakura.tomoya}@jp.fujitsu.com

satoshi.sekine@riken.jp inui@ecei.tohoku.ac.jp

1 はじめに

企業が運営する製品やサービスのサポートセンターでは、頻繁に問い合わせのある質問と回答のペアをFAQとして準備しており、コールセンターのオペレータの回答のための知識源としての利用したり、ユーザの検索対象として Web 上で提供されている。

ユーザによるFAQ検索では入力された自然文やキーワードを基に検索が行われるが、その問題の一つとして、FAQの質問と回答に含まれる語彙と、ユーザが検索の際に使う語彙が異なるという点が挙げられる。たとえば、ユーザが「ホームページが閲覧できません」と検索した場合、パソコン関係のFAQにおいて「インターネットの接続の問題がある場合」といった回答の候補が存在していたとしても、異なる語彙で構成されていれば、そのページを検索することは難しい。この問題への対処方法の一つとしては、ユーザからの質問とそれに対応するFAQのペアを利用した機械学習によるモデル構築が考えられる。しかし、コールセンター運用開始直後や、新規のFAQにおいては、ユーザからの問い合わせが関連付けられたFAQが存在しないため、機械学習の適用が行えない。

本稿では、ユーザが用いる語彙による質問を自動生成し、回答検索器の作成を行う手法を提案する。図1に提案手法の概要を示す。まず、「①質問生成器学習」として、ユーザが使う語彙から構成される質問を生成するために、ユーザがオンラインでやりとりを行っているQA集を用いて、質問生成器を学習する。続いて、「②質問生成」で、①の質問生成器を用いて、検索対象のQA集の各QAに対して、新たな質問を複数付与する。「③回答検索器学習」では、②で付与された質問を用いて、検索モデルを学習する。最後に、「④回答検索」にて、ユーザからの質問を用いて、検索を行う。オンラインQA集として、Yahoo!知恵袋の質問と回

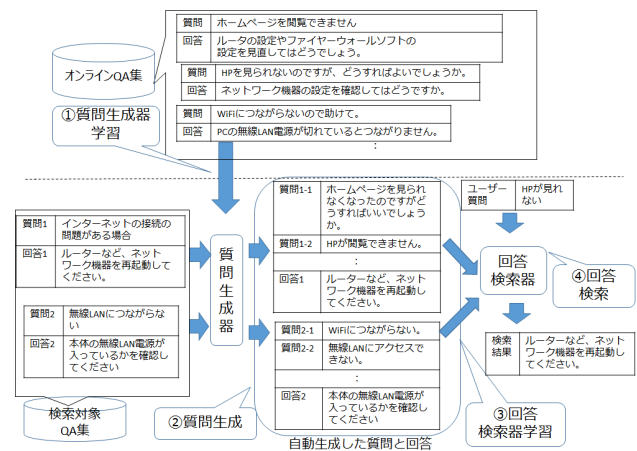


図 1: 提案手法概要

答のペアを用い、Twitter 上の富士通の PC 関連のサポートに関する対応履歴から作成した FAQ 検索用のデータにて、評価した。実験結果から、FAQ および、Yahoo!知恵袋から学習した検索モデルと比較し、自動生成した質問を用いて学習した検索モデルが、高い回答精度が得られることが示された。

2 質問生成器利用の利点

目的のタスクの学習データがない場合における対処方法の一つとして、Yahoo!知恵袋などのコミュニティ型質問応答サイトの質問と回答から学習データを生成する方法が考えられる。この方法と比較し、質問生成器に用いた学習データ生成は次の利点が考えられる。

1. 検索対象に対する学習。提案手法は生成したクエリを用いるため、実際の検索対象に対する学習が行える。一方、コミュニティ型質問応答サイトの質問と回答を検索の学習に用いる場合は、実際とは異なる検索対象における学習となる。

2. 非適合文書の選択．今回の事例である FAQ 検索では，FAQ は通常，重複が無いように作成されることが一般的であり，学習時の非適合文書選択の問題が軽減される．一方，コミュニティ型質問応答サイトを検索の学習に用いる場合は，コミュニティ型 QA サイトは類似の QA ペアが存在することもあり，正しい回答となり得る事例を負例として選択しうる．

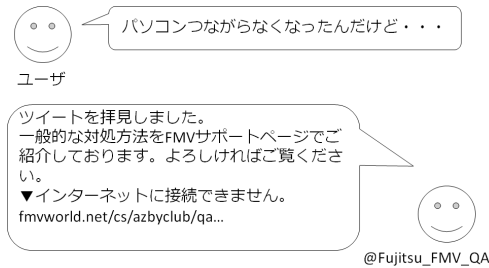


図 2: @Fujitsu.FMV_QA のリプライ

3 提案手法

提案手法は質問生成器の学習，回答検索のための学習データ生成，回答検索学習のステップからなる．

3.1 質問生成器の学習

質問生成器の学習には，機械翻訳の分野で用いられている，sequence to sequence [4] の枠組のうち，注意付き RNN エンコーダ・デコーダを用いて学習した [2]．質問生成器の学習データには，Yahoo!知恵袋を用いた．¹ Yahoo!知恵袋はスレッドからなり，スレッドはタイトル，質問本文および回答のリストから構成される．また，回答リストの中から，ベストアンサーが選出されている．本稿では質問本文の単語列 $\mathbf{y} = \langle y_1, \dots, y_M \rangle$ およびベストアンサーの単語列 $\mathbf{x} = \langle x_1, \dots, x_N \rangle$ の対から学習データを生成する．Yahoo!知恵袋データから生成した学習データを $D = \{(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_n, \mathbf{y}_n)\}$ とする．学習では負の対数尤度を最小化するようパラメータ ϕ を更新する．

$$L(\phi) = \sum_{(\mathbf{x}, \mathbf{y}) \in D} -\log p(\mathbf{y}|\mathbf{x}; \phi).$$

3.2 回答検索のための学習データ生成

本稿では FAQ を検索対象とし，FAQ を 1 つの回答とみなす． i 番目の FAQ d_i は，質問および回答 a_i からなるものとする．3.1 節で学習した質問生成器に，FAQ の回答 a_i を入力として，ビームサーチによって n 件の質問リスト \mathbf{h}_i を生成する．生成された質問に対して，その FAQ d_i は適合文書とする．

3.3 回答検索学習

検索ではモデルによって与えられるスコアが最大となる FAQ を返す．

$$d = \operatorname{argmax}_{d' \in S} F_\theta(q, d'),$$

ただし S は検索対象の FAQ の集合とし， $F_\theta(q, d)$ をパラメータを θ としたときの質問 q に対する FAQ d のスコアとする．

学習時は，質問とその適合文書のペアの集合が $R = \{(q^{(1)}, d^{(1)}), \dots, (q^{(n)}, d^{(n)})\}$ と与えられた際に，

$$\begin{aligned} \theta &= \operatorname{argmin}_{\theta'} E(\theta'), \\ E(\theta') &= \sum_{(q^{(i)}, d^{(i)}) \in R} \left\{ \max(0, l_{\theta'}(q^{(i)}, d^{(i)})) \right\}, \\ l_{\theta'}(q^{(i)}, d^{(i)}) &= \max_{d' \in S(q^{(i)})} F_{\theta'}(q^{(i)}, d') - F_{\theta'}(q^{(i)}, d^{(i)}), \end{aligned}$$

を最小にするような θ を求めることになる．ただし $S(q^{(i)})$ は質問 $q^{(i)}$ に対して検索対象となる FAQ の集合とする． $E(\theta')$ は検索に関するエラーとなり，全ての質問に対して，正しい回答を返すことができれば，値は，0 となる．

本稿では検索モデルとして Supervised Semantic Index (SSI) [1] をランキング向けに改良した版を用いる [3]．SSI では質問側と，検索対象側の単語分散表現をそれぞれ学習して，以下の式にて，スコアを設定する．

$$F_{\mathbf{U}, \mathbf{V}}(\mathbf{q}, \mathbf{a}) = \mathbf{q}^\top (\mathbf{U}^\top \mathbf{V} + \mathbf{I}) \mathbf{a}.$$

\mathbf{q} と \mathbf{a} が質問と回答に出現する単語集合を意味するベクトル， \mathbf{U} と \mathbf{V} が質問と回答の単語に対応する単語分散表現であり， \mathbf{I} は単位行列とする．

4 実験

質問生成器を用いて生成した学習データの良さを評価するために，以下の学習データを用いてランキング

¹http://www.nii.ac.jp/dsc/idr/yahoo/chiebrkr2/Y_chiebukuro.html のデータを利用した．

モデルを学習する．表 1 に，各データのサイズを載せる．学習データとしては，以下を用いる．

FAQ 富士通の FAQ．富士通のパソコンサポートサイトで公開されている 13,129 件の FAQ の各タイトルを質問とし，その FAQ の回答部分を適合文書とする．非適合文書は，各 FAQ 以外の 13,128 件とする．

Y! Yahoo!知恵袋データの質問とベストアンサー「>インターネット,PCと家電>パソコン」>インターネット,PCと家電>ソフトウェア」および「>インターネット,PCと家電>インターネット」に含まれるスレッドを対象から抽出した．各質問のベストアンサーが適合文書となる．各質問と適合文書に対して，5 件の非適合文書を同カテゴリ内のベストアンサーから無作為に抽出する．

QGEN (n) 質問生成器を基に FAQ から自動生成した質問と各 FAQ．富士通の FAQ 13,129 件のそれぞれに n 件の質問を付与するので， $n \times 13,129$ 件の質問と適合文書が得られる．各質問と適合文書の非適合文書としては，FAQ の集合から，それぞれ 5 件を，無作為に抽出する．n としては，10 と 50 を設定した．

Tw @Fujitsu_FMV_QA のユーザへのリプライ．@Fujitsu_FMV_QA のリプライにあるユーザからの質問とその適合文書の対から得られた学習データ 4,918 件を用いる．非適合文書は，以下に記載する評価時の検索対象と同じ 1,134 件のうち，適合文書を除いた 1,133 件とした．このデータは，実際のユーザのクエリが得られる場合との比較を行うために作成した．また，評価データと重複しないように選択している．

評価データは，Twitter 上でのアカウント @Fujitsu_FMV_QA と他の Twitter ユーザとのリプライから作成した．@Fujitsu_FMV_QA は Twitter 上のユーザからの問い合わせに回答したり，ストリーム上の Tweet からパソコン関係で発言しているユーザを見つけて，解決手段を回答している．その際，@Fujitsu_FMV_QA は富士通のパソコンサポートサイト²上で公開されている FAQ などを引用して回答を行っている．本稿では，@Fujitsu_FMV_QA の引用した FAQ を適合文書，Twitter ユーザの tweet を質問として抽出する．図 2 の例では「パソコンつながらなくなったんだけど・・・」が質問となり，@Fujitsu_FMV_QA

²<http://azby.fmworld.net/support/>

表 1: データサイズ．QGEN(10)，QGEN(50) が提案手法による学習データ．

データ	(質問, 適合文書) 数
FAQ	13,129
Y!	399,292
QGEN (10)	131,280
QGEN (50)	656,400
Tw	4,918
評価データ	1,213

が参照した URL のタイトル「インターネットに接続できません」を適合文書とする．評価に使う回答候補の FAQ としては，収集した @Fujitsu_FMV_QA のツイート中で引用された 1,134 件を利用する．

評価ではユーザの各質問において，回答として提示された 1 件の適合文書，残りの 1,133 件を非適合文書として，各質問を与えて，1,134 件の FAQ をランキングした結果を基に，評価値を算出する．評価値には Mean Reciprocal Rank (MRR)，Recall@1 および Recall@10 を用いる．

ベースラインとして tfidf に基づくランキングを評価する．次の様にして質問に対して FAQ のスコアを計算し，スコアの降順に FAQ を並べ変えて出力する:

$$f(q, d) = q \cdot d,$$

$$q_i = \frac{tfidf(q_i, q, C_q)}{\sum_{j=1}^{|q|} tfidf(q_j, q, C_q)},$$

ただし $tfidf(q_i, q, C_q)$ は q における単語 q_i の頻度，FAQ の質問部分からなるコーパス C_q における q_i の文書頻度に基づいて計算される tfidf 値とする． d_i は d および FAQ 回答の回答部分からなるコーパス C_d から同様に計算する．

5 実験結果

表 2 に実験結果を載せる．ユーザの質問が学習時に利用できない場合は，提案手法である自動生成した質問を用いた QGEN (50) が，MRR で最も高い精度となり，平均的に正解が上位に来ていることがわかる．R@10 では他の手法より精度が低くなっているが，R@1 でも最も高い精度が得られている．また，ユーザの質問が含まれる Tw を用いて学習した結果が最も高い精度であることから，本手法にて，目的の検索対象に関して，ユーザの語彙による質問が生成され，精度が改善したと予想される．

表 2: 実験結果

手法	学習データ	学習時質問	学習時検索対象	MRR	R@1	R@10
tfidf	-	-	-	0.144	0.088	0.269
SSI	FAQ	FAQ のタイトル	FAQ の回答部分	0.142	0.084	0.267
SSI	Y!	質問	ベストアンサー	0.166	0.097	0.302
SSI	QGEN (10)	生成された質問	FAQ のタイトル	0.164	0.103	0.261
SSI	QGEN (50)	生成された質問	FAQ のタイトル	0.178	0.133	0.265
SSI	Tw	ユーザのツイート	FAQ のタイトル	0.357	0.282	0.513

表 3: 提案手法の成功例

質問	学習データ	各モデルで得られた 1 位の FAQ
パソコンから音がでなくなった。 困った。	Y!	[アップデートナビ] 困ったときの対処方法を教えてください。
	QGEN (50)	スピーカーから音が出ません。

表 3 に提案手法の成功例を示す。提案手法がこの事例で成功している要因として、この FAQ の回答から「パソコンを立ち上げた時に音が出ません。どうしたらいいですか?」や「pc を立ち上げた時に音が出なくなってしまう。どうしたらいいのでしょうか?」のような質問を生成していることが挙げられる。

6 関連研究

質問を生成する方法として、文献 [6] の手法では対応ログが蓄積されている状態で、各 FAQ に関連するログを選別し、それらを使って学習する方法を提案している。学習用のデータを生成する点では同一であるが、今回の方法は、そのようなログが蓄積されていない状態で、質問を生成して学習する方法である。

また、Wieting ら [5] は質問の翻訳結果を利用する方法を提案している。一方、本手法は、関連する分野に絞って、質問生成器を学習し、質問を生成する点で異なり、本手法と組合せた利用も可能である。

7 おわりに

本稿は FAQ 検索モデルを学習するための学習データを Yahoo!知恵袋から生成する方法を提案した。質問生成器に基づいて生成された学習データによって学習した検索モデルは、Yahoo!知恵袋のみを使って学習した場合よりも高い MRR および Recall@1 を示した。今後の課題として多様なバリエーションの質問を生成するための質問生成方法を検討したい。

参考文献

- [1] Bing Bai, Jason Weston, David Grangier, Roman Collobert, Kunihiko Sadamasa, Yanjun Qi, Olivier Chapelle, and Kilian Weinberger. Supervised Semantic Indexing. In *Proc. of CIKM'09*, 2009.
- [2] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective Approaches to Attention-based Neural Machine Translation. In *Proc. of EMNLP'15*, pp. 1412–1421, 2015.
- [3] Takuya Makino and Tomoya Iwakura. A boosted supervised semantic indexing for reranking. In *AIRS'17*, pp. 16–28, 2017.
- [4] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Proc. of NIPS'14*, pp. 3104–3112, 2014.
- [5] John Wieting, Jonathan Mallinson, and Kevin Gimpel. Learning paraphrastic sentence embeddings from back-translated bitext. In *Proc. of EMNLP'17*, pp. 274–285.
- [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*, Vol. 24, No. 1, pp. 117–134, 2017.