トピックモデルに基づく多様性を考慮した回答文書検索に関する研究

情報数理応用研究　　5218C006-5　大川順也
指導教員　後藤正幸

A Study on Answering Documents Retrieval Considering Diversity Based on Topic Models

OKAWA Junya

1. 研究背景・目的

近年，自然言語で記述されたユーザの質問に対して，自動的に回答を行う質問応答システム (Question Answering(QA) Systems) の利用場面は増えている．一般的な質問応答システムでは，ユーザから与えられた質問を解析し，Web ページや新聞記事などの情報源となる文書集合から，質問に対する適切な回答を検索するモデル (回答文書検索モデル) を構築することで回答候補を抽出し，回答の自動化を実現している [1].

一方，こうした質問・回答文書データが集まるサイトの 1 つに，コミュニティ QA(Question Answering) サイト (cQA サイト) と呼ばれる質問回答サイトがある．cQA サイトは，ある利用者が匿名もしくはユーザ ID を使用することで質問を投稿し，別の利用者が質問に対する回答を投稿することで成り立っている．このような cQA サイトにおける質問・回答文書に対して，回答文書検索モデルを構築する場合，基本的な手法として質問文書の類似度による手法が考えられる [2]．この手法は，新しく対応したい質問文書と類似する質問文書を過去の質問履歴から検索し，類似質問文書と対応している回答文書を提示するという手法である．

しかしここで，一般に cQA サイトにおける利用者の年齢・性別・職業・利用目的などは様々であり，多種多様な質問が投稿されている．また，質問によって求められる回答も多種多様である．例えば,「一般常識」や「科学的事実」についての質問には，適切な回答が存在するものが多くある．一方で,「人生相談」についての質問には，多様な解が存在し，望ましい回答は質問者によって異なる．この時，検索モデルはこの回答文書の多様性を適切に検索結果に反映させる必要がある．質問文書の類似度による手法は，質問文書の内容のみを考慮し検索・提示を行っている．そのため，cQA サイトで見られるような，質問ごとに存在する回答文書の多様性を的確に掴みながら，適切な内容の回答文書を提示することは困難であると考えられる．

そこで本研究では，トピックモデルの代表的な手法である Latent Dirichlet Allocation(LDA)[3] を用いることで，従来手法では対応できなかった回答文書の多様性を考慮可能な回答文書検索モデルの構築手法を提案する．ここで，筆者ら [4] はすでに，質問文書と回答文書の関係性を，LDA を用いることで，トピックによって定量的に分析する手法を提案している．この手法を援用することで，過去の質問・回答履歴を学習し，質問文書と回答文書の関係性をトピックによって紐付ける．これにより，新しく対応したい質問文書に対する回答文書として，適切なトピックを推定することができ，回答文書との類似度に基づいて直接適切な回答文書を検索することが可能となる．以上により，提案手法では，質問のトピックだけではなく，回答のトピックも考慮した，新しい検索モデルの枠組みが構築される．加えて，回答文書の多様性を的確に掴みながら，適切な内容の回答文書を検索・提示することが可能になると期待できる．提案モデルの有効性を検証するため，実際に cQA サイトに投稿された質問・回答文書を用いた検証実験を行い，検索結果の性能を評価するとともに，得られた結果に基づいて考察を行う．

2. 準備

2.1. 質問応答システム

ユーザからの質問に対して自動で適切な回答を提示するシステムを質問応答システムといい, 多くの研究が行われている．例えば，Zhang ら [5] は，双方向の Long Short-Term Memory(BiLSTM) に基づいた協調処理による回答選択手法を示した．一般的な質問応答システムは，新規に与えられた質問文書を解析し，類似文書を情報源から検索することで回答候補を抽出している．対して提案モデルは，過去の質問・回答事例が与えられたもとで，質問文書の類似度は考慮せず，新規の質問に対する適切な回答の特徴表現を推定することで，類似回答文書を検索・提示するという特徴がある．

一方で，我々が研究対象としている cQA サイトにおける質問・回答文書を用いた研究も近年盛んに行われている．例えば，横山ら [6] は，新規に投稿された質問に対して適切な回答を提示することが可能な回答者を探索する手法を提案している．また，西原ら [7] は質問文書と複数の回答文書が与えられたもとで，ベストアンサーになる可能性が高い回答を判定する手法を提案している．その際，質問者と回答者の相性に着目し，質問と回答の文末表現から相性を評価することでベストアンサーの判別を行っている．このように cQA サイトを対象として，適切な回答者の推薦，ベストアンサーの判別などの研究が行われており，本研究では cQA サイトに寄せられる新規の質問に対して，適切な回答を同サイト上で投稿された過去の回答履歴から検索するという問題設定を扱う．

2.2. Latent Dirichlet Allocation(LDA)

不特定多数の人物が自由に作成している質問・回答文書には，話題や内容，利用者ごとの書き方の特徴など多様な情報が混在している．このような多様な情報が混在した文書の生成過程をモデル化するための有効な手法として LDA[3] が知られている．

いま，文書集合を $\mathcal{D} = \{1, \cdots, N\}$，単語集合を $\mathcal{V} =$

$\{1,\cdots,V\}$，トピック集合を $\mathcal{K}=\{1,\cdots,K\}$ と表記する．文書 $d\in\mathcal{D}$ がトピック $k\in\mathcal{K}$ に所属する確率を $\theta_{d,k}$ としたとき，文書 d のトピックへの所属確率を表す分布 (トピック分布) を $\boldsymbol{\theta}_d=(\theta_{d,1},\cdots,\theta_{d,K})$ と表記する．また，トピック k の下で単語 $v\in\mathcal{V}$ が出現する確率を $\phi_{k,v}$ とし，トピック k の単語の出現確率の分布（単語分布）を $\boldsymbol{\phi}_k=(\phi_{k,1},\cdots,\phi_{k,V})$ と表記する．文書ごとのトピック分布 $\boldsymbol{\theta}_d$ とトピックごとの単語分布 $\boldsymbol{\phi}_k$ に対しては，それぞれハイパーパラメータ $\boldsymbol{\alpha}=(\alpha_1,\cdots,\alpha_K)$,$\boldsymbol{\beta}=(\beta_1,\cdots,\beta_V)$ を持つディリクレ分布を事前分布として仮定する．このとき LDA のモデル式は式 (1) で表される．

$$p(v|d)=\sum_{k=1}^{K}\iint\theta_{d,k}\phi_{k,v}P(\theta_{d,k}|\alpha_k)P(\phi_{k,v}|\beta_v)d\theta_{d,k}d\phi_{k,v} \tag{1}$$

3. 提案手法

3.1. 概要

従来の検索手法では，質問文書の類似度に基づいて回答文書を提示している．しかしながら，cQA サイトなどでは，質問によって，多様な回答文書が存在している．従来手法では，質問文書の内容のみを考慮し検索をしているため，回答文書の多様性を的確に掴みながら検索・提示することが困難であると考えられる．そこで本研究では，回答文書の多様性に対応可能な，新しい枠組みの検索モデルの構築手法を提案する．提案モデルでは，質問・回答双方の内容を考慮し，回答文書のトピックの類似度に基づいて直接的に適切な回答文書を検索・提示する．これにより提案モデルでは，従来手法では対応困難な回答文書の多様性を考慮しながら，適切なトピックの回答文書を提示できると期待できる．

具体的には，提案手法では，まず初めに筆者らがすでに提案したトピック関連度スコア算出手法 [4] を用いることで，質問・回答文書の関係性をトピックにより定量的に表現する．次に，トピック間の関係性をもとに，新しく対応したい質問文書に対する適切な回答文書の特徴表現を算出する．最後に適切な回答文書の特徴表現と類似する回答文書を検索・提示する．

3.2. トピック関連度スコアの算出

本節では，筆者らがこれまでに提案したトピック関連度スコアの算出手法について説明する．トピック関連度スコアとは，質問文書と回答文書の関係性を，トピックによってスコア付けしたものである．

以降の説明では，質問・回答文書に適用する LDA をそれぞれ Q-LDA，A-LDA，質問・回答文書のトピックをそれぞれ Q-Topic，A-Topic と定義する．提案モデルでは，はじめに，Q-LDA，A-LDA により質問・回答文書のトピック分布 $\boldsymbol{\theta}_d$,$\boldsymbol{\theta}'_d$ を推定する．

次に，質問・回答文書のトピック間の関連度が高いときは，それぞれのトピックへの所属確率の積も高くなるという仮定のもと，式 (2) によりトピック関連度スコアを算出する．ここで，$\gamma^d_{k,l}$ を d 番目の質問・回答文書対における，Q-Topic k と A-Topic l 間のトピック関連度スコアとし，$\boldsymbol{\Gamma}^d$ は $\gamma^d_{k,l}$ を (k,l) 要素とする行列である．$\gamma^d_{k,l}$ は Q-LDA，A-LDA により推定した d 番目の質問・回答文書対の Q-Topic k と A-Topic l への所属確率の積により算出する．また，$\boldsymbol{\Gamma}$ は最終的に求まるトピック関連度スコア行列を表し，全ての文書対のトピック関連度スコア行列を足し合わせることで算出する．$\gamma_{k,l}$ を最終的な質問・回答文書対における，Q-Topic k と A-Topic l 間のトピック関連度スコアとすると，$\boldsymbol{\Gamma}$ は $\gamma_{k,l}$ を (k,l) 要素とする行列となる．

$$\boldsymbol{\Gamma}=\sum_{d=1}^{N}\boldsymbol{\Gamma}^d\qquad(\boldsymbol{\Gamma}^d=\boldsymbol{\theta}_d^{\mathrm{T}}\cdot\boldsymbol{\theta}'_d) \tag{2}$$

最後に以下の式 (3) により，Q-Topic ごとにトピック関連度スコアの和が 1 になるように正規化する．このようにトピック関連度スコアを正規化することで，Q-Topic ごとの，それぞれの A-Topic に対する関連度合いを割合で把握することが可能となる．

$$\widetilde{\gamma}_{k,l}=\frac{\gamma_{k,l}}{\sum_j\gamma_{k,j}} \tag{3}$$

3.3. 回答文書検索モデル

提案モデルでは，過去の質問・回答のトピック関連度スコアを用いることで，新しく対応したい質問文書に対する回答文書として適切なトピック分布を推定し，推定された A-Topic 分布に基づき検索を行う．以降の説明では，この過去の質問・回答のトピック関連度スコアが与えられた時に，新規の質問文書のトピック分布に対する回答文書のトピック分布として適当だと推定できる分布を最適 A-Topic 分布と定義する．最適 A-Topic 分布を導入することにより，従来の質問文書のみの類似度による検索ではなく，回答文書との類似度に基づき直接適切な回答文書を検索することが可能なモデルを構築することができる．

提案モデルでは，まず 3.2 節に従って質問・回答文書のトピック関連度スコアを算出する．次に新しく対応したい質問文書に対する最適 A-Topic 分布を算出する．ここで，回答文書のトピック数を L，d 番目の質問文書に対する最適 A-Topic 分布を $\boldsymbol{\theta}^*_d=(\theta^*_{d,1},\cdots,\theta^*_{d,L})$ とすると，l 番目の要素 $\theta^*_{d,l}$(A-Topic l への所属確率) はトピック関連度スコアと新規の質問文書に対する Q-Topic 分布を用いて以下の式 (4) に従って算出する．

$$\theta^*_{d,l}=\sum_{k=1}^{K}\theta_{d,k}\widetilde{\gamma}_{k,l} \tag{4}$$

最後に，最適 A-Topic 分布と類似する A-Topic 分布を持つ回答文書を過去の回答履歴から検索する．このとき分布間の類似度の計算には KL 情報量を用いる．

以上より，回答文書検索手順は以下のようになる．

Step1) LDA を適用することで，全ての質問・回答文書に対してトピック分布を準備

Step2) 前節で述べたトピック関連度スコアの算出手法を用いて，トピック関連度スコアを学習

Step3) 式 (4) により，新しく与えられた質問文書に対する最適 A-Topic 分布を算出

Step4) 最適 A-Topic 分布との距離が KL 情報量の意味で近い A-Topic 分布を持つ過去の回答文書を検索し提示

4. 評価実験

4.1. 実験条件

本実験では，cQA サイトである「Yahoo!知恵袋」に実際に投稿された質問・回答文書データを使用する．Yahoo!知恵袋では，質問に対する「ベストアンサー」が，質問者の選択，もしくは他の利用者からの投票によって選ばれる．本実験では 2014 年 1 月から 2017 年 3 月に投稿された質問文書と質問に対するベストアンサーをデータセットとして用いる．さらに，Yahoo!知恵袋では，質問はカテゴリごとに分けられ，質問者が質問投稿時にカテゴリを選択するようになっている．本実験では 7 つのカテゴリ「音楽」「映画」「生き方，人生相談」「政治，社会問題」「ニュース，事件」「天気，天文，宇宙」「資格，習い事」に属する質問とその質問に対応する回答を実験の対象とした．本実験では，100,672 件のデータセットの中から，それぞれのカテゴリに属する文書を無作為に 30 件ずつ抽出しテストデータとして，それ以外のデータを学習データとして用いる．それぞれのテストデータに対する過去の回答の適合判定は，人手により行った．判定基準は，質問に対する回答として，適切なトピックであり，違和感が無い内容である回答を正例，それ以外を負例としている．

本実験の比較手法として，従来手法である質問文書の類似度による手法を用いる．質問文書のベクトル化には BoW と Q-Topic 分布の 2 つを用いる．BoW を用いた手法では cos 類似度を，Q-Topic 分布を用いた手法では KL 情報量を用いて文書同士の類似度を測る．

また，本研究では，各文書の単語頻度ベクトルを構成するために，質問文書と回答文書それぞれに対して形態素解析を行い，名詞，動詞，形容詞の単語を抽出した．単語頻度ベクトルの次元数は $V = 4,263$ である．また，LDA を適用する際のトピック数は，質問・回答文書のトピック数 K, L={100,200,300,400,500} と変更させて実験を行った．検索精度の評価指標としては，MAP(Mean Average Precision) を用いる．MAP は精度と再現率の両方を重視した指標であり，情報検索の評価指標として広く用いられている．本実験では，それぞれの手法で類似度上位 3 件を検索し，このときの MAP を計算することで検索精度を評価する．

4.2. 実験結果

4.2.1. 検索精度に関する分析

各手法による検索精度の結果を表 1 に示す．表 1 における BoW，Q-Topic 分布はそれぞれ前節で述べた BoW と Q-Topic 分布を用いた質問文書の類似度による手法を示している．なお提案モデルの結果は，検索精度上位 5 組の K と L の組合せの時の MAP を示している．

表 1 より，提案モデルが従来手法より高い MAP を示していることが分かる．これにより，質問・回答文書のトピック関連度を考慮した検索により，比較的適切な内容の回答文書を検索できることが分かる．

表 1: 各手法の検索精度

モデル		MAP
BoW		0.460
Q-Topic 分布	K=100	0.524
	K=200	0.545
	K=300	0.550
	K=400	0.558
	K=500	0.563
提案	$K = 500, L = 400$	0.584
	$K = 500, L = 300$	0.576
	$K = 500, L = 500$	0.573
	$K = 500, L = 200$	0.568
	$K = 400, L = 400$	0.566

また表 2 に，各カテゴリの提案モデルと Q-Topic 分布を用いた手法の MAP を示す．なお表 2 において，それぞれ最も高い MAP を示した提案モデルは質問・回答文書のトピック数 $K = 500, L = 400$ の時の結果を，Q-Topic 分布を用いた手法は $K = 500$ の時の結果を示している．

表 2: 各カテゴリの提案モデル ($K = 500, L = 400$) と Q-Topic 分布を用いた手法 ($K = 500$) の検索精度

カテゴリ	提案	Q-Topic 分布
資格，習い事	0.475	0.621
天気，天文，宇宙	0.510	0.639
映画	0.578	0.583
音楽	0.591	0.592
政治，社会問題	0.634	0.497
ニュース，事件	0.639	0.516
生き方，人生相談	0.658	0.493

表 2 より，質問のカテゴリに応じて 2 つの手法の検索精度に違いがあることが分かる．たとえば，「政治，社会問題」，「ニュース，事件」，「生き方，人生相談」に関する質問に対しては，提案モデルの検索精度の方が明らかに高い．一方，「資格，習い事」，および「天気，天文，宇宙」に関する質問に対しては，Q-Topic 分布を用いた手法の検索精度の方が明らかに高いことがわかる．このようにカテゴリによって検索精度に差が現れるのは，質問のカテゴリによって回答文書の多様性に違いがあり，2 つの手法でこの回答文書の多様性の掴み具合が異なっているためと考えられる．次節で，従来手法と提案モデルの実際の検索結果から，双方の回答文書の多様性の掴み具合の差異を分析する．

4.2.2. 検索結果に関する分析

本節では，2 つの質問に対する，Q-Topic 分布を用いた手法と提案モデルの実際の検索結果の分析を行う．1 つ目として，「人生相談」についての質問である Q1「人間が成長するのにはどんな事をするのがいいと思いますか？」に対する 2 つの手法の検索結果を表 3，4 に示す．表 3，4 のラベルは，回答の適合判定の結果を示している．このような質問に対しては，多様な回答が存在していると考えられる．したがって検索モデルはこの回答文書の多様性を適切に検索結果に反映させる必要がある．提案モデルの検索結果を見ると，どれも「人間の成長」についての多様な回答となっていることが分かる．対して，表 3 を見ると，Q-Topic 分布を用いた手法は，違うトピックの回答を検索していることも分かる．

表 3: Q1 に対する Q-Topic 分布を用いた手法の検索結果

Rank	検索回答文書	ラベル
1	努力は財産です． 一番困るのは，何にもない人です． 何も出来ない．何処にも行けない．	○
2	近くをうろついて，ぶつかってあげれば	×
3	芸術として，単純性を良く表現して 良いのではないでしょうか	×

表 4: Q1 に対する提案モデルの検索結果

Rank	検索回答文書	ラベル
1	何でもチャレンジ！ やってみなければ解らない	○
2	幸せになりたいからです．自問自答 試行錯誤を繰り返して，人は成長していく のだと思います	○
3	生きていれば何にしろ経験します． それらの経験で人格形成，協調性，情緒， 向上心等が生まれ人として　行きます．	○

次に，2 つ目の質問 Q2「どうして空は青いのですか？」に対する 2 つの手法の検索結果を表 5，6 に示す．Q2 は知識に対する質問である．この場合，回答は限定的であり，多様な回答も存在していない．表 5 より，Q-Topic 分布を用いた手法は全ての結果で，質問に対する的確な回答文書を検索している．一方で，表 6 より，提案モデルは Q2 に対しては求められている回答を的確に検索できていないことが分かる．すなわち，提案モデルは勉学系などの単一の回答が求められている質問に対しては有用な結果を示さないと考えられる．しかし，人生相談，人間関係の悩みなど多様な回答が求められている質問に対して，ユーザにとって有用な検索結果を示すと考えられる．

表 5: Q2 に対する Q-Topic 分布を用いた手法の検索結果

Rank	検索回答文書	ラベル
1	空が青く見えるのは，太陽光の可視光領域 のうち，波長の短い青色の光を 大気中の気体分子が散乱させ，その 散乱された光 (青色の光) を見ているからです．	○
2	可視光線の中の，青や紫色光線の 気体散乱現象によるという話です．	○
3	太陽の光が空気中のある粒子に反射， あるいは屈折し，散乱しやすい青が 空の色となっています．	○

5. 考察

本研究における主な着眼点は，提案モデルにより新規に与えられた質問文書に対する適切な回答文書を提示できるか，また，cQA サイトで見られるような質問ごとに存在する回答文書の多様性をうまく掴むことができるのかということであった．実験結果より，提案モデルでは従来手法より比較的高い検索精度を示した．従来手法は質問文書の特徴量のみに基づき検索しているのに対し，提案モデルは，質問・回答のトピックの関係性を考慮し，過去の回答履歴から適切なトピック分布をもつ回答文書を直接検索する．これにより，提案モデルはより適切な内容を含む回答文書を検索することが可能となり，比較的高い検索精度を示したと考えられる．また提案モデルの検索精度に関して，$K = 500, L = 400$ の場合に MAP が最大となっている．質問文書は文書長が長く単語数も多いため，トピック数 K を大きくすると，より詳細な文書表現が得られ検索精度が高くなったと考えられる．一方で，

表 6: Q2 に対する提案モデルの検索結果

Rank	検索回答文書	ラベル
1	雲の中は湿度 100 %です．雲に近づくほど 一般に湿度は高くなります．	×
2	気象庁では，風速の単位は秒速がわかりやすい と考え，長期にわたって秒速で表示しています．	×
3	進化論でいえば海や空が赤く見える動物は絶滅， 青く見える動物だけ生き残った，と思います．	×

質問文書に比べ回答文書は文書長が短いため，$L = 500$ に比べ $L = 400$ の場合に，より適切な文書表現を得られ検索精度が高くなったと考えられる．

また実際の検索結果の分析より，提案モデルは多様な回答が求められている質問に対して，有用な結果を示した．これは，新規の質問に対する適切な回答のトピック分布を用いて検索することにより，ある程度揺らぎを持たせながら回答を直接検索することが可能となり，多様な回答文書を提示できたためと考えられる．一方で，一般教養や勉学系のようなピンポイントな回答が求められている質問に対しては，過去にも同じ内容の質問が存在するような場合が多く，従来手法によりその質問を検索することができるため，従来手法と提案モデルは適切に組み合わせて使い分けることが有用であると考えられる．

6. まとめと今後の課題

本研究では，質問・回答文書双方のトピックの関係性を考慮した新しい枠組みの回答文書検索モデルの構築手法を提案した．実データを用いた実験の結果より，提案モデルが検索精度を保ちつつ，多様な回答が求められている質問に対して有用な検索結果を与えることを示した．

今後の課題としては LDA におけるトピック数の最適な決定が挙げられる．現在は，実験的にトピック数を決めている．したがってトピック数を文書データの情報などから決定することが今後の課題として挙げられる．

参考文献

[1] Suzuki, J., Sasaki, Y., and Maeda, E., “SVM Answer Selection for Open-Domain Question Answering,” *Proceedings International Conference on Computational Linguistics*, pp.974–980,2002.

[2] C.D.Manning., P.Raghavan., and H.Schuetze., Introduction to Information Retrieval, Cambridge University Press, 2008.

[3] Blei M.D., Ng Y.A., and Jordan I.M., “Latent Dirichlet Allocation,” *Journal of Machine Learning Research,* vol.3, pp.993–1022, 2003.

[4] 大川順也, 雲居玄道, 後藤正幸, “潜在的ディリクレ配分法を用いた問合せ文書と回答文書の関係分析,” 経営情報学会誌, (掲載決定), 2020.

[5] Zhang, L., and Ma, L., “Coattention based BiLSTM for Answer Selection,” *Proceedings of the 2017 IEEE International Conference on Information and Automation(ICIA)*, pp.1005–1011, 2017.

[6] 横山友也, 宝珍輝尚, 野宮浩揮, “質問回答サイトにおける質問文への適切な回答者の選出法,” 日本感性工学会論文誌, 第 15 巻, 第 1 号, pp.21–29, 2016.

[7] 西原陽子, 松村真宏, 谷内田正彦, “QA サイトにおける質問に適した回答の判定,” NLP 若手の会 第 2 回シンポジウム, 2007.