多様な質問を受け付ける質問応答システムの回答選択手法

横川　裕太　　白井　清昭
北陸先端科学技術大学院大学 情報科学研究科
{s1310075, kshirai}@jaist.ac.jp

1 はじめに

質問応答システムに関する初期の研究は，事実を回答とするファクトイド型の質問応答が主流であったが，近年は理由や方法など，様々な質問に答えることのできる質問応答システムへの期待が高まっている．ユーザからの多様な質問に答えるためには，質問文からユーザがどのような情報 (理由，方法，人物など) を求めているかを特定する必要がある．回答となる情報の種類による質問文の分類は回答タイプ (または質問タイプ) と呼ばれている．一般的な質問応答システムは，回答タイプをあらかじめ定義し，質問文のタイプを判別した上で，知識源となる文書集合から得られた複数の回答候補の中からそのタイプに適合する回答を選択する．しかし，多様な質問を受け付ける質問応答システムでは，回答タイプをあらかじめ網羅的に定義することは難しい．事前に定義されていないタイプの質問が入力されると，回答候補を適切に絞り込むことができないという問題が生じる．

本研究では，明確な回答タイプの分類を行わず，多様な質問を受け付けることのできる質問応答システムの一手法について述べる [1]．提案システムでは，回答候補の適切さを「内容の関連度」と「回答タイプの整合度」で評価する．内容の関連度は，質問と回答候補の内容やトピックの類似性を測る．一方，回答タイプの整合度は，質問と回答候補の回答タイプが一致しているかを評価する．例えば，理由を尋ねる質問に対し，何らかの理由を説明しているテキストは回答タイプが一致するが，用語の定義を説明しているテキストは回答タイプが一致しないとする．ただし，回答タイプの存在を仮定しているが，回答タイプはあらかじめ定義せず，単に暗黙の回答タイプが一致しているかを評価している点に特徴がある．本論文では，特に (1) 上記 2 つの観点に基づくスコアの組み合わせ方，(2) 回答タイプの整合性を判定する二値分類器を学習するための訓練データの作成方法について論じる．

2 提案手法

2.1 システムの概要

提案システムにおける処理の流れを図 1 に示す．通常の質問応答システムと同様に，質問文解析，文書検索，回答抽出という流れで処理を行う．質問文解析モジュールでは，質問文から自立語を抽出して検索クエリを作成すると同時に，後述する回答タイプの整合度を算出するための素性を抽出する．文書検索モジュールでは，ウェブ検索エンジン [1] を用いて文書検索を行い，上位 N 件 (本論文では $N = 100$) の HTML 文書を獲得する．得られた HTML 文書を div や p タグなどでパラグラフに分割し，それらを回答候補とする．回答抽出モジュールでは，まず回答候補のスコアを計算する．質問と回答候補の内容の関連度 (2.2 項で詳述) と，回答タイプの整合度 (2.3 項で詳述) を算出し，両者から 2.4 項で述べる方法で最終的な回答候補のスコアを計算する．最後に，スコアの上位の候補を回答として出力する．

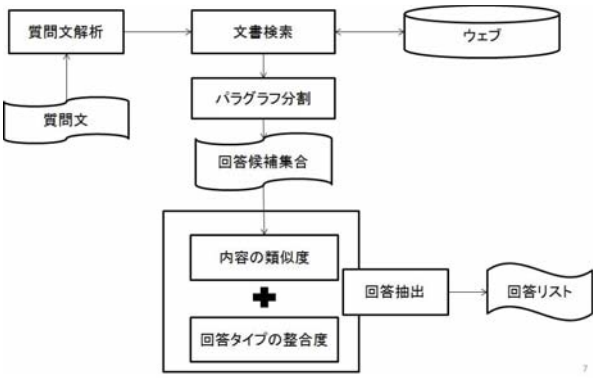

図 1: 質問応答システムの概要

2.2 内容の関連度

質問と回答候補の内容の関連度スコア $Score_r$ は，質問文，回答候補に含まれる自立語から構成される単語ベクトルをそれぞれ $\vec{q}$, $\vec{a}$ とし，それらのコサイン類似度 $cos(\vec{q}, \vec{a})$ により求める．さらに，回答候補抽出の精度を高めるために以下の処理を行う．

質問文中の重要な単語に高い重みを与える
質問文の単語ベクトル $\vec{q}$ において，助詞「が」「は」「の」の前にある語の重みを 2，それ以外の語の重みは 1 とする．

文書中の前後のパラグラフを考慮
回答候補の単語ベクトル $\vec{a}$ において，その回答候補の前後のパラグラフ中の自立語を単語ベクトルに加える．その重みは 0.5 とする．

[1] 実験では Bing の検索 API (http://datamarket.azure.com/dataset/bing/search) を利用した．

ウェブ検索結果のランキング

ウェブ検索エンジンによる検索結果の上位の文書に含まれる回答候補は質問との関連が高いと考え，ウェブ検索時のランクが高いほどスコアが高くなるように補正する．

$Score_r$ の定義を式 (1) に示す．r は回答候補を含む文書のウェブ検索におけるランク，N は回答候補を抽出する文書数である．

$$Score_r = cos(\vec{q}, \vec{a}) \times (1 - \frac{r}{N}) \quad (1)$$

2.3 回答タイプの整合度

質問と回答候補の組が与えられたとき，それらの回答タイプが一致しているかを判定する二値分類器を学習する．学習アルゴリズムは L2 正則化ロジスティック回帰を用いた．使用したツールは LIBLINEAR [2] である．LIBLINEAR が出力する確率値 (回答タイプが一致する確率) を回答タイプの整合度 $Score_t$ とする．以下，回答タイプの一致を判定する分類器を学習する手法について述べる．

2.3.1 素性

機械学習に用いる素性を以下に示す．

$\mathbf{f_{in}}$ 質問文に現れる疑問表現 (「何」「誰」など)．

$\mathbf{f_{in3}}$ 疑問表現を含む 3-gram．ただし，自立語は品詞に置き換える．

$\mathbf{f_{end}}$ 質問文の文末に現れる単語列 (文末表現と呼ぶ)．

$\mathbf{f_{cl}}$ 回答候補の節 [3] の末尾に現れる単語列 (節末表現と呼ぶ)．この素性は文末表現も含む．

$\mathbf{f_{end+cl}}$ 回答候補における文末表現と節末表現の組み合わせ．

$\mathbf{f_{func}}$ 回答候補に出現する付属語の列．

ここで，$\mathbf{f_{end}}$, $\mathbf{f_{cl}}$, $\mathbf{f_{cl+end}}$ における文末表現ならびに節末表現とは，以下のいずれかである．

- ⟨回答タイプ指示語⟩+付属語の列
- ⟨動詞⟩+付属語の列 [4]
- ⟨自立語の品詞⟩+付属語の列

⟨回答タイプ指示語⟩ とは,「方法」「原因」「理由」「対処」など，回答タイプを特定する手がかりとなると考えられる 40 個のキーワードである．回答タイプ指示語のリストは人手で作成した．図 2 に素性の例を示す．

2.3.2 訓練データの作成

分類器を学習するための訓練データは，Yahoo!知恵袋 [5] における 76,782 個の質問と回答 (ベストアンサー) の組から作成する．Yahoo!知恵袋では質問の内容に応じて 14 のカテゴリー (趣味，コンピュータなど) が設定されている．

質問文: そばとうどんの違いは何ですか?
回答候補: 蕎麦はそば粉で作られますが，うどんは小麦粉から作られます．

$\mathbf{f_{in}}$ 何
$\mathbf{f_{in3}}$ ⟨名詞⟩_は_何, は_何_です, 何_です_か
$\mathbf{f_{end}}$ 何ですか
$\mathbf{f_{cl}}$ 作られますが, 作られます
$\mathbf{f_{end+cl}}$ 作られますが+作られます
$\mathbf{f_{func}}$ は, で, れ_ます_が, から, れ_ます

図 2: 素性の例

正例 (回答タイプが一致している事例) は，Yahoo!知恵袋における質問と回答の組とする．一方，負例は，質問と，その質問とは別の質問に対する回答の組から作成する．ただし，回答タイプが同じ質問と回答の組を誤って負例とすることを避けるため，元の質問と似ていない質問に対応する回答を選ぶ．訓練データ作成の手続きを以下に示す．

1. Yahoo!知恵袋から質問と回答の組 (q, a) を選び，これを正例とする．
2. Yahoo!知恵袋における他の質問から，回答タイプの一致という観点で q との類似度の小さい質問 q'_i を N_{neg} 個選択する．質問間の類似度は，2.3.1 で述べた質問文の素性 ($\mathbf{f_{in}}$, $\mathbf{f_{in3}}$, $\mathbf{f_{end}}$) から成る素性ベクトルのコサイン類似度で測る．
3. 元の質問 q と q'_i に対応する回答 a'_i の組 (q, a'_i) を負例とする．
4. 1〜3 の操作を全ての質問について繰り返す．

次に，N_{neg} の決め方，すなわち訓練データにおける正例と負例の比の決め方について述べる．正例と負例の比は，質問応答システムの文書検索モジュールによって得られる回答候補の集合における正例と負例の比に近いことが望ましい．そこで，本研究では，図 3 に示す文書検索のシミュレーションによって N_{neg} を決める．まず，$\mathbf{f_{in}}$, $\mathbf{f_{in3}}$, $\mathbf{f_{end}}$ から成る素性ベクトルを用いて，Yahoo!知恵袋における質問と回答の組をクラスタリングする．クラスタリングツールとして CLUTO[6] を利用し，クラスタリングアルゴリズムは Repeated Bisection 法を選んだ．クラスタ数は 15 に設定した．ここでは，同一クラスタに属する質問は回答タイプが同じであるとみなす．次に，質問を検索クエリとして回答集合に対して文書検索を行い，関連度 (質問と回答に含まれる自立語から成る単語ベクトルのコサイン類似度) の高い上位 N 件の回答を取り出す．ここで N は質問応答システムにおけ

[2] http://www.csie.ntu.edu.tw/~cjlin/liblinear/
[3]「節」は文を接続助詞で区切った断片と定義する．
[4] このパターンは $\mathbf{f_{end}}$ の抽出には用いない．
[5] http://chiebukuro.yahoo.co.jp/
[6] http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview

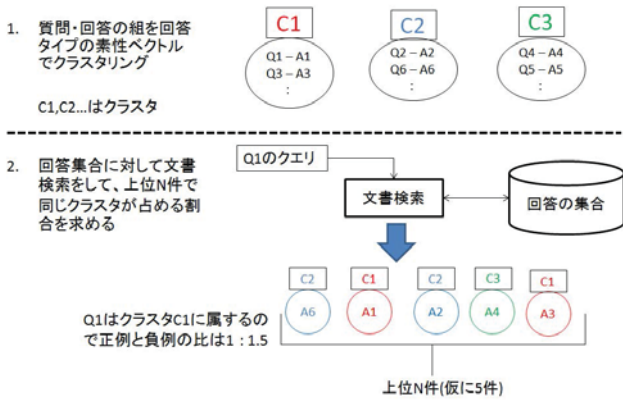

図 3: 正例と負例の比の求め方

る文書検索モジュールが取得する文書数と同じ値 (本論文では $N = 100$) に設定する．得られた回答のうち，クエリの質問と同じクラスタに属する (≒暗黙的に同じ回答タイプである) 回答を正例，それ以外を負例とみなし，正例と負例の比を求める．上記の処理は，Yahoo!知恵袋のデータをカテゴリ毎にサブコーパスに分割し，そのサブコーパス毎に行う．全ての質問について求めた比の平均を最終的な正例と負例の比とする．

2.4 回答候補のスコア

回答候補のスコアは前述の「内容の関連度」と「回答タイプの整合度」という 2 つの観点から決める．本研究では以下の 2 通りの方法を提案する．

フィルタリング方式

回答タイプの一致を判定する分類器をフィルタとして利用する．この方式では，回答タイプが一致しないと判定された回答候補を除外する．残りの回答候補のスコアは式 (1) に示した $Score_r$ とする．

加算方式

$Score_r$ と $Score_t$ の重み付き和により回答候補のスコアを決定する．ここで，2 つのスコアの重みをどのように最適化するかが問題となる．開発データを用意して最適化することも考えられるが，多様な質問を含む開発データを用意することは現実的には難しい．本研究では式 (2) のように両者を組み合わせる．

$$Score = Rel\text{-}Score_r \times 0.5 + Rel\text{-}Score_t \times 0.5 \quad (2)$$

ここで $Rel\text{-}Score_r$，$Rel\text{-}Score_t$ はそれぞれ $Score_r$，$Score_t$ の相対スコアであり，回答候補の集合における最大のスコアに対する比と定義する．

3 評価実験

3.1 手順

まず，評価用の質問セットを用意した．提案システムは様々な種類の質問を受け付けることを目標にしているため，実験では 7 種類の質問を設定した．表 1 に質問の種類とその定義を示す．次に，それぞれについて 5 個，計 35 個の質問を用意した．質問はインターネット上の雑学サイトなどから収集した．

表 1: 評価に用いた質問の分類

比較	～と…の違いは何?
事実確認	～はどうなっていますか? ～は本当ですか?
ファクトイド	～した人物は誰? ～がある場所はどこ?
定義	～とはなんですか? ～はどのような人物ですか?
方法	～はどうやりますか? ～はどうすればいいですか?
理由	なぜ～なのですか?
意見・感想	～についてどう思いますか? ～はどうでしたか?

評価用の質問セットに対して提案システムが出力した回答集合を以下の 3 つの評価基準で評価した．MRR(Mean Reciprocal Rank) は，最上位の正答のランクの逆数の平均であり，式 (3) のように定義される．Q は評価に用いた質問の集合であり，$rank_k$ は質問 k における正答の最高順位である．

$$MRR = \frac{1}{|Q|} \sum_{k \in Q} \frac{1}{rank_k} \quad (3)$$

AP' は，質問応答システムによって出力された上位 10 件の回答の平均精度 (average precision) であり，式 (4) のように定義される．N_k は質問 k に対する出力上位 10 件中の正解数であり，P_i は i 番目の正答を出力した時点での精度を表す．

$$AP' = \frac{1}{|Q|} \sum_{k \in Q} \left(\frac{1}{N_k} \sum_{i=1}^{N_k} P_i \right) \quad (4)$$

P_{top10} は，質問応答システムが回答を 10 件出力したときの精度であり，式 (5) のように定義される．

$$P_{top10} = \frac{1}{|Q|} \sum_{k \in Q} \frac{N_k}{10} \quad (5)$$

3.2 結果と考察

実験の結果を表 2，3，4 に示す．これらの表において，「フィルタリング方式」と「加算方式」は 2.4 項で述べた回答候補の順位付け手法を表わす．さらに，回答タイプの一致を判定する分類器の学習データにおける正例と負例の比を 1:1，1:5.9，1:10 としたときの結果を比較した．ここで，1:5.9 は 2.3 項の手法によって決定された正例と負例の比である．また，比較のため，回答タイプの整合度のスコアを用いず，内容の関連度のスコア $Score_r$ のみで回答選択したときをベースラインとし，その結果を「BL」の列に示す．

内容の関連度と回答タイプの整合度の両方を考慮したフィルタリング方式や加算方式は，ベースラインよりも評価値が高いことから，回答選択において回答タイプの一致を考慮することは重要であると言える．また，2 つの方式を比較すると，加算方式の方がフィルタリング

表 2: MRR による評価

	フィルタリング方式			加算方式			BL
	1:1	1:5.9	1:10	1:1	1:5.9	1:10	
比較	0.70	0.73	0.36	0.90	0.67	0.65	0.70
事実	0.71	0.55	0.60	0.65	0.81	0.72	0.57
ファ	0.54	0.48	0.32	0.43	0.41	0.80	0.53
定義	0.36	0.67	0.16	0.42	0.54	0.57	0.40
方法	0.28	0.61	0.35	0.62	0.63	0.49	0.21
理由	0.32	0.25	0.41	0.14	0.43	0.47	0.39
意見	0.38	0.60	0.65	0.46	0.90	0.55	0.21
全体	0.47	0.55	0.41	0.52	0.63	0.61	0.43

表 3: AP' による評価

	フィルタリング方式			加算方式			BL
	1:1	1:5.9	1:10	1:1	1:5.9	1:10	
比較	0.65	0.56	0.35	0.75	0.62	0.47	0.64
事実	0.61	0.44	0.46	0.50	0.65	0.60	0.51
ファ	0.54	0.39	0.25	0.42	0.39	0.70	0.53
定義	0.33	0.60	0.18	0.38	0.48	0.43	0.39
方法	0.36	0.54	0.31	0.49	0.54	0.40	0.26
理由	0.33	0.27	0.41	0.12	0.30	0.31	0.32
意見	0.29	0.57	0.54	0.45	0.75	0.42	0.22
全体	0.44	0.48	0.36	0.44	0.53	0.48	0.41

表 4: P_{top10} による評価

	フィルタリング方式			加算方式			BL
	1:1	1:5.9	1:10	1:1	1:5.9	1:10	
比較	0.34	0.30	0.18	0.36	0.36	0.32	0.32
事実	0.16	0.12	0.12	0.22	0.14	0.20	0.18
ファ	0.16	0.14	0.16	0.22	0.22	0.26	0.18
定義	0.16	0.34	0.16	0.24	0.26	0.30	0.26
方法	0.24	0.28	0.12	0.26	0.28	0.22	0.20
理由	0.18	0.20	0.16	0.08	0.22	0.16	0.24
意見	0.24	0.38	0.24	0.36	0.44	0.24	0.22
全体	0.21	0.25	0.16	0.25	0.27	0.24	0.22

方式よりも全般に結果が良かった．これは，回答タイプの一致を判定する分類器の正解率[7]が十分に高くなく，フィルタリング方式では分類器によって不一致と判定した回答候補を除外することで正答を誤って取り除く場合が多いためと推察される．

正例と負例の比については，1:5.9 が 1:1 や 1:10 と比べて評価値が高いことから，提案手法による正例と負例の比の決定方法の有効性が確認できた．また，1:5.9 の結果を質問の種類ごとに見てみると，定義やファクトイドの質問ではフィルタリング方式のほうが MRR や AP' が高いことがわかる．ファクトイドや定義型のいわゆる「何」を問う質問は訓練データによく出現することから，回答タイプの判別が比較的容易で，フィルタリング方式が効果的に働くためと考えられる．

4 関連研究

水野らは Q&A サイトの質問・回答事例を学習データとし，質問と回答のタイプが一致しているかを判定する SVM を学習する手法を提案している [2]．この研究でも回答タイプを事前に定義せず，暗黙の回答タイプの一致を判定している．また，Q&A サイトに投稿された質問と回答の組を正例，質問とその質問とは別の回答の組を負例としている．ただし，水野らの手法では正例と負例の割合を単に 1:1 に設定しているのに対し，本研究では正例と負例の比の最適化を試みている．また，負例となる質問と回答の組を選択する際に，質問文の素性 ($\mathbf{f_{in}}$, $\mathbf{f_{in3}}$, $\mathbf{f_{end}}$)，すなわち回答タイプが似ていない事例を選ぶことによって，回答タイプが一致する質問と回答の組を誤って負例としない工夫を試みた点が異なる．

石下らは，Q&A サイトのデータ集合から回答の記述スタイルを表す特徴的な単語 n-gram を動的に抽出し，これを基に「タイプの整合度」のスコアを算出し，またこのスコアと「内容の関連度」のスコアの重み積によって回答候補のスコア付けを行う [3]．しかし，特徴的な単語 n-gram を動的に求めるため，回答候補選択に要する計算時間が大きいという問題点がある．また 2 つのスコアの重みを変えた実験結果を示しているが，重みの最適化までは行っていない．本研究でも重みの最適化は行っていないが，上記 2 種のスコアを組み合わせる 2 つの手法を提案し，実験的に比較した．

Soricut らは，統計的機械翻訳の技術を応用し，質問文から回答文への翻訳のスコアをもとに回答選択を行う手法を提案している [4]．この手法も回答タイプをあらかじめ定義していないが,「内容の関連度」と「タイプの整合度」の両方を同時に学習するため，大量の訓練データを必要とするという問題点がある．

5 おわりに

本論文では，多様な質問に答えられる質問応答システムにおいて，質問と回答の内容の類似度，及び両者の回答タイプの整合度を基に適切な回答を選択する手法を提案した．回答タイプの一致の判定精度を向上させるために，現在提案している素性の組み合わせや新しい素性を探究することが今後の課題である．

参考文献

[1] 横川裕太: 多様な質問を受け付ける質問応答システムの回答選択に関する研究,修士論文, 北陸先端科学技術大学院大学,2015.

[2] 水野 淳太, 秋葉 友良: 任意の回答を対象とする質問応答のための実世界質問の分析と回答タイプ判定法の検討, 言語処理学会 13 回年次大会発表論文集,pp.1002-1005,2007.

[3] 石下 円香, 佐藤 充, 森 辰則: Web 文書を対象とした質問の型によらない質問応答手法, 人工知能学会論文誌 24 巻 4 号 B,pp.339-350,2009.

[4] Soricut,R. and Brill,E.: Automatic Question Answering Using the Web:Beyond the Factoid, Journal of Information Retrieval Special Issue on Web Information Retrieval, Vol.9,pp.191-206,2006.

[7] 訓練データの 10 分割交差検定での正解率は 85.79%であった．