質問応答システムにおける利用者の回答選択支援のための出典抽出手法の提案

櫻井 大平†　渋木 英潔‡　森 辰則‡

†横浜国立大学 大学院 環境情報学府　‡横浜国立大学 大学院 環境情報研究院
E-mail: {d_saku,shib,mori}@forest.eis.ynu.ac.jp

1　はじめに

Web 上に存在する情報は，ブロードバンド化の進展やブログ等の普及に伴い，爆発的に増加し続けている．そこで，質問応答などの技術を用いた，必要とする情報に効率良くアクセスする方法が望まれている．従来の質問応答システムは，Web 文書などの情報源から複数の回答候補（情報源に現れる語句やパッセージ）を抽出し，抽出された回答候補群において，独自の基準によりスコアを計算しランキングする方法が一般的であった．しかしながら，必ずしも利用者の求める回答が上位にランキングされる訳ではなく，正解と不正解の回答が混在した状態で利用者に提示されているのが現状である．そのため，利用者は正解と不正解の回答が混在した状態で，正解となる回答を判断しなければならない．

この問題に対し，我々は，図 1 に示すように，回答ごとにその出典となる情報を示すことができれば，利用者の回答判断を支援できると考えた．例えば，図 1 の場合，回答 3 は出典が存在しないため，回答 1 や回答 2 と比べて正しさが疑わしいと判断することができる．また，回答 1 と回答 2 では，回答 2 の出典がより新しいため，かつては回答 1 が正解と考えられていたが，現在は回答 2 が正解と考えられているのではないかと判断することができる．このような背景の下，我々は，対象となる回答の出典情報を提示することで，情報の信頼性を判断する基準を提供し，利用者の回答選択を支援することを目的とした研究を行う．

本研究における出典とは，図 2 に示すように，回答の抽出元である元文書ではなく，その回答の元文書が参照している，情報の出所となる文書等を意味している．したがって，出典には新聞記事や書籍など様々なものが考えられるが，本研究で前提とする質問応答システムが Web 文書を情報源とすることから，出典も Web 上で参照可能な文書に限定することとした．また，そのような文書への参照は一般に URL が用いられていることから，本稿では回答の元文書から URL により参照されている出典の Web 文書を**出典文書**と定義することとした．

最終的には，図 1 に示すように，出典文書の URL（以降，**出典 URL** と呼ぶ）に加えて，出典文書の内容を端的に示す主題や年度などを提供することが目標であるが，本稿ではその第一段階として出典 URL の抽出を対象とした手法を提案する．提案手法は，ある回答とその元文書が与えられた時に，元文書に存在する URL 群の中から，与えられた回答の出典を示すものとして適切な URL を抽出するものである．ここで，元文書に存在する URL は出典文書を参照するものだけではないことに注意されたい．また，文書中の記述の出典文書を示す URL であっても，回答とは別の記述に関する URL である場合も考えられる．そのため評価実験の節で説明するように，回答記述に最寄りの URL を採用するといった素朴な手法では F 値が 0.3 程度であり，元文書に存在する URL 群の中から，回答の出典を示す適切な出典 URL を見つけることは決して容易な課題ではない．本稿では，2 節で関連研究を述べた後，3 節で，元文書に存在する URL に対して，回答の出典 URL として適切であるかの判定に関する議論を行い，それに基づいた，各 URL の出典 URL としての適切性を判定する手法を 4 節で述べる．

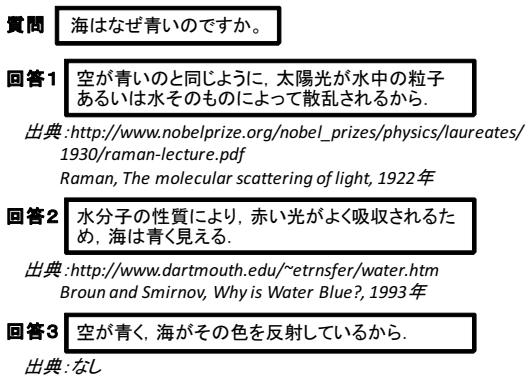

図 1: 出典による回答選択支援の例

2　関連研究

URL を抽出する研究には，北村らの研究 [1] がある．北村らは Wikipedia を対象として，出典・注釈情報から自動的に出典を分類する手法を提案している．北村らが，対象とする Web 文書の形式を Wikipedia に限定しているのに対し，本研究では質問応答の情報源として用いられる，形式の定まっていない Web 文書を対象としている点が異なる．

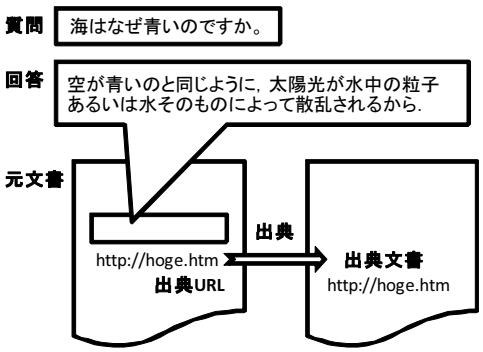

図 2: 回答と出典の関係

3 元文書中の URL の出典 URL としての適切性判定の枠組み

適切性判定の議論を行うに先立ち，回答候補，URL，関連文書を以下のように定義する．質問応答システムの回答候補の各々「回答候補$_i$」について，回答候補の抽出元文書からとりだされた，各URL_{ij}について，まず，＜回答候補$_i$, URL_{ij}＞の組をつくり，この組に対してURL_{ij}の出典 URL としての適切性の判断を行う事を考える．しかし，URL_{ij}がテキストではなく記号列であることから，この組が持つ情報だけでは，出典 URL としての適切性判断は難しい．そこで手掛りとして，組＜回答候補$_i$, URL_{ij}＞に「関連」する文章を見つけてきて追加し，＜回答候補$_i$, URL_{ij}, 関連文章$_{ijk}$＞の組に対し，URL_{ij}が回答候補$_i$の出典 URL として適切かどうかを判断する事を考えた．この関連文章をどのような「関連性」に基づき採取するのかについて，以下の (A),(B) に示す二通りの方法を検討した．

(A) 当該 URL による参照関係に基づく関連性

一つ目は，＜回答候補$_i$, URL_{ij}, 関連文章$_{ijk}$＞の組において，関連文章として，URL が指し示す先の文書，即ち実際の情報が書いてある出典文書を設定する．この場合，URL 一つに対して曖昧性なく関連文章が一つだけ得られるため，実際には，＜回答候補$_i$, URL_{ij}, URL_{ij}で指示される文書＞の組が適切性の判定対象の組となる．この時，URL_{ij}で指示される文書の内容自身が質問応答の回答に相応しいものであれば，出典である可能性が高いと考えた．つまり，URL_{ij}で指示される文書の内容が，元々の質問応答の回答とし適切な記述を含むか否かで評価する．

(B) 当該 URL が現れる別文書の内容に基づく関連性

二つ目は，関連文章として，当該 URL を含有する，回答の元文書とは別の文書（同一 URL 含有文書），より正確には，同一 URL 含有文書において当該 URL が現れる周囲の文脈を扱う．この場合，そういった文脈の選択は多くの可能性があるため，ある一つの文脈に注目するだけでは判断できず，一種の多数決処理を行うこととなる．すなわち，数多くの文脈において当該 URL が現れ，その文脈群が回答候補と密接な関連があるのであれば，その URL は有効な出典候補を指し示している，という仮説に基づき適切性の判定を行う．

そういった関連文章の取得の方向性には二通りあり，＜回答候補, URL, 関連文章＞の組において，URL と回答候補のどちらを起点にして関連文章を取得し，類似した文脈の数の多さを調査するかという点が異なる．

(B-1) URL を起点とする場合

＜回答候補, URL, *＞として，回答候補と当該 URL を固定し，関連文章すなわち当該 URL の出現する文脈を求める．すなわち URL を固定して，それが現れる文書群を集め，その文書群の各文脈に対し，回答候補との類似性を判断することにより，＜回答候補, URL, *＞の組の出典としての適切性を判断する事を行う．

(B-2) 回答候補を起点とする場合

＜回答候補, URL, *＞における回答候補に類似する文書群を取得する．すなわち，＜回答候補, *1, *2 ＞として，まず*2 の候補を集め，その文書中に現れる URL を*1 とする．その上で，*1 に当該 URL が出現する頻度によって，＜回答候補, URL, *＞の組の出典としての適切性を判断する事を行う．

以上の事から，適切性判定における観点は以下の 3 通りに整理することができる．
A) 出典文書の内容が質問応答の回答に相応しいかどうか．
B-1) 当該 URL を起点として取得した，同一 URL 含有文書における内容の類似性が高いかどうか．
B-2) 回答候補を起点として取得した文書中の URL における，当該 URL の頻度が高いかどうか．

4 リンク構造を利用した出典 URL の抽出

4.1 出典 URL の抽出処理の概要

目的とする出典 URL の抽出に対し，1) 出典 URL 候補の抽出，2) 前節で議論した適切性の判定，の二段階からなる手法を提案する．図 3 にその概要を示す．

(1) 質問応答システムに質問を入力し，回答を得る．ここまでは既存の質問応答システムを用いる．

(2) 各回答に対しその回答の元文書を Web から収集する．

(3) (2) で取得した各回答に対する元文書に対し，それぞれ元文書に現れる URL を出典 URL の候補として抽出する．

(4) (3) で出力された各 URL に対し，出典 URL としての適切性の判定を行う．

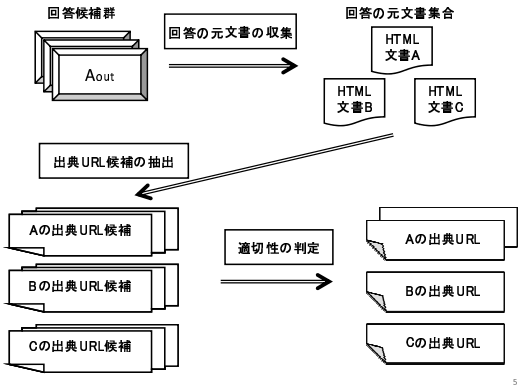

図 3: 出典 URL の抽出処理の流れ

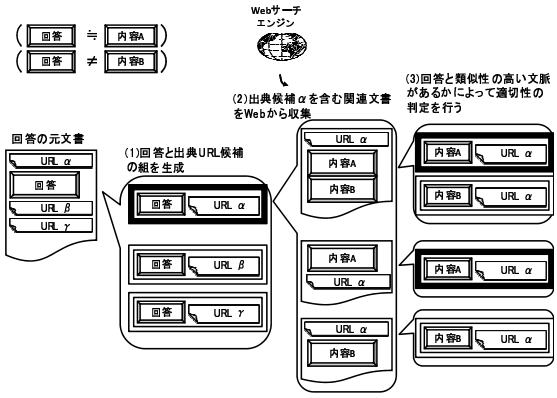

図 4: URL を起点とした判定手法

4.2 出典 URL 候補の抽出

実際の Web 文書における出典の現れ方を踏まえ，回答の周囲，回答の元文書の上部・下部に記載されている URL を出典 URL 候補とした．ここで，回答の元文書には，回答と類似する内容及びその内容に対する出典が書かれている場合もあるため，元文書に現れる URL 全てを出典 URL 候補とすると，後の適切性判定を経たとしても，適切な出典 URL の抽出がうまくいかないと考えた．そのため，HTML における文書構造を表すタグを利用して，出典 URL 候補を絞り込む事を行う．一方で，出典が文書の冒頭や末尾にまとめられることも多いので，元文書の上部と下部は出典 URL 候補を探索する範囲とする．

(1) 回答の周囲の探索

回答として抽出された箇所から，まず上方向に探索し，見出しを表す <hN>（N は 1~ 6 の数字）か水平の罫線を表す <hr> が出現するまでに記載されている URL を出典 URL 候補とする．次に上方向と同様に下方向にも探索するが，上方向で <hN> により探索を終了していた場合，階層構造を考慮するため，当該 <hN> よりも下位の階層であるタグ <hM>（M>N）では探索を終了しない．

(2) 回答の元文書の上部・下部の探索

上部については，回答の元文書において <body> の直後に記載されている URL の列について，いずれも出典 URL 候補とし，下部については </body> タグの直前に記載されている URL の列について，同様に出典 URL 候補とする．

各探索において，出典候補とする URL の最大数は，(1),(2) それぞれ n 件までとし，実験では n=5 とした．

4.3 出典 URL 候補の適切性の判定

A 参照関係を利用した適切性判定

当該 URL が指し示す文書に対する，質問応答の回答としての適切性を判定する上で，質問応答システムによって予め収集されている回答の元文書集合を利用することを考える．この元文書集合の各々に対し，質問応答システムが何らかの回答候補が存在すると判断している．よって，当該 URL 先の文書が元の質問に対する回答の元文書集合に含まれているならば，回答の出典として適切である，といった仮説に基づき適切性の判定を行った．以下に具体的な処理の流れを示す．

(1) 質問応答システムから出力された回答の元文書の URL リストを作成する

(2) 出典 URL 候補として抽出された URL が (1) で作成したリストに存在する場合，それを出典 URL とし，存在しなければ出典 URL ではないと判定する

B-1 URL を起点として共起関係を利用

図 4 にその概要を示し，処理の流れについて説明する．

(1) 回答が抽出された元文書において，回答と出典 URL 候補の組を生成する．

(2) 各出典 URL 候補である URL に対し，対象 URL を含む文書を Google Custom Search API[1] を用いて Web から収集し関連文章とする．図 4 では，出典 URL 候補 α に注目し，α を含む文書を Web から収集している．

(3) (2) で収集した各文書に対し，当該 URL の現れる文脈と回答候補との類似度を計算し，複数の文脈において回答との類似度が高くなった場合，関連文書検索に用いた出典 URL 候補 α は出典として適切だと判定する．

処理の流れ (3) において，文脈の切り出しは <p>, といった段落を表すタグ情報を手掛かりとして行った．以下の実験では，類似度は，各テキストを，そのテキスト中の形態素 2-gram の有無を成分とするベクトルに対応付け，cosine 類似度を用いて計算し，類似度が高いと判断する閾値は 0.1 とした．

B-2 回答候補を起点として共起関係を利用

[1]https://developers.google.com/custom-search/?hl=ja

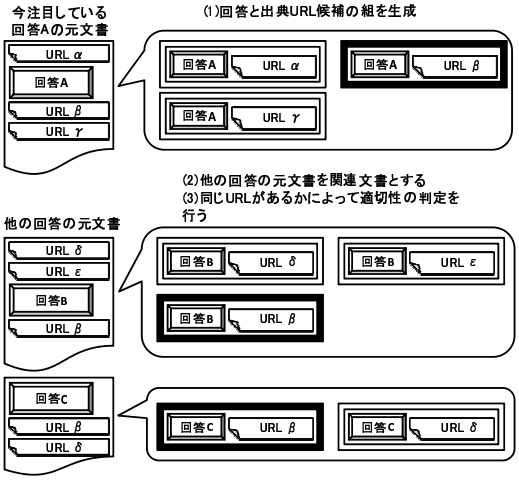

図 5: 回答候補を起点とした判定手法

図 5 にその概要を示し，処理の流れについて説明する．

(1) 回答が抽出された元文書において，回答と出典 URL 候補の組を生成する．

(2) 今注目している回答に類似する文書集合を関連文章とする．以下の実験では，質問応答システムが出力する回答候補集合には類似性が存在すると仮定して，他の回答の元文書集合を関連文章とした．

(3) (2) で収集した文書集合において，(1) における出典 URL 候補の出現頻度を求める．閾値以上の頻度を持つものを出典 URL として適切であると判断する．以下の実験では閾値を 2 とした．その際に，当該 URL が異なるサイトの文書に現れる頻度を重視するために，以下の 2 点の条件を定めた．

・URL のサーバ名が同一の文書から抽出された URL は，同じ文書から抽出された URL とみなす．
・元文書の URL と抽出された URL のサーバ名が同一である場合，その URL は対象に含めない．

5 評価実験・考察

提案手法の有効性を調べるために，4 章の 3 つの処理，及び 3 つの処理の中で 1 つ以上の処理で出典と判定されたものを出力する処理（和集合）について評価実験を行った．

5.1 評価方法

実験には「どのような状態を人の「脳死」と呼びますか。」等，評価型ワークショップ NTCIR-6 の QAC-4 タスクセットにおける 3 つの質問に対し，質問応答システムの出力した回答上位 100 件を用いた．回答に対する出典 URL として，適切な URL を正解情報とし，一つの回答に対し複数の出典 URL があった場合は，そのいずれかが抽出された時，正解が抽出されたと見なした．正解情報については，4.2 節で述べた探索範囲において人手で判断して作成し，次式の適合率・再現率・F 値を用いて評価を行った．

適合率=抽出された正解数/システムの抽出結果
再現率=正解の抽出できた回答数/正解のある回答数
F 値　=(2 ×適合率×再現率)/(適合率＋再現率)

また比較のため,「回答候補が抽出された箇所に最寄りの URL を出典 URL と判断する」といったベースライン手法を設けた．その結果を表 1 に示す．

表 1: 評価実験結果

	適合率	再現率	F 値
ベースライン	0.23 (27/116)	0.39 (27/69)	0.29
(A) 参照	0.71 (5/ 7)	0.07 (5/69)	0.13
(B-1) 共起 (URL)	0.84 (32/38)	0.46 (32/69)	0.59
(B-2) 共起 (回答)	**0.94** (17/18)	0.25 (17/69)	0.39
(A)+(B-1)+(B-2) 和集合	0.84 (49/58)	**0.71** (49/69)	**0.77**

5.2 考察

個別の手法においては，手法 (B-1) が F 値において 0.59 と最も高くなっている．これは，他の 2 手法と異なり Web 検索を用いた関連文章の拡張を行った事で，再現率が高くなったためと考えられる．手法 (A) については，限定された状況を扱ったため，出力数に限界があったと考えられる．手法 (B-2) については，回答の元文書の上部・下部に記載されている出典 URL が抽出されやすい傾向にあった．これは文書全体の出典として現れる出典 URL が優位になったためと考えられる．また 3 手法の単純な和集合を求めた結果，再現率・F 値において個別の手法より良い結果となった．これは，各手法により抽出された出典の重複が少なく，結果一定の精度を保ちつつ再現率が上昇できたためと考えられる．

6 おわりに

質問応答システムの回答に対する出典を抽出する手法の検討を行った．出典としての適切性を判定するための 3 手法を提案し，各手法の出力の和を求める手法により，評価実験において F 値 0.77 の結果を得た．関連文書の拡張手法の拡充による適切性判断の精緻化が今後の課題である．

参考文献

[1] 北村 大樹, 山田 剛一, 絹川 博之: “Wikipedia 出典・脚注情報の媒体分類の自動付与”, 情報科学技術フォーラム講演論文集, **9**(2), pp.303-304, (2010).