Web 文書の抜粋を回答とする質問応答システム
Question Answering with Web Page Chunks

山本 正範　　延澤 志保　　太原 育夫
東京理科大学 理工学部 情報科学科

1　まえがき

Web 上には膨大な情報が蓄積されている．Web 情報は幅広く，そして最新の情報が存在するため，情報源として有用である．Web から情報を取得する手段としては検索エンジンを用いることが一般的であるが，Web 上のページを検索し，表示するだけでなく，Web に存在する膨大な情報を知的に活用しようとする試みが行われてきている．その一つに，Web を情報源とした質問応答システムの研究がある．質問応答システムの多くは，質問に対して名詞句で答えるシステムであり，文章で答えることのできるシステムは少ない．これはシステムが文章を生成することが極めて困難なためである．本研究では，Web 上に存在する膨大な情報に着目し，自然言語文での質問に対して回答として適切な文章を Web から抽出することで，回答文生成処理なしで提示することができるシステムを実装し，評価を行った．

2　Web を情報源とする質問応答システム

質問応答システムの情報源として，特定のコーパスや事典，新聞記事などを用いることが多いが，情報量の豊富な Web を情報源とする質問応答システムの研究がなされている．

2.1　システム構成

Web を情報源とた質問応答システムの構成は以下のようになる [1][2][3].

質問解析部
ユーザから受けた質問文を解析し，検索エンジンの形に合った検索式を生成する．また，質問の種類を決定する．

検索部
検索エンジンを用いて Web から文書を取得する．

回答抽出部
収集した文書から，回答の候補となる部分を抽出する．

回答選択部
回答の候補を選別し，順位をつけてユーザに返す．

2.2　既存研究

Web を情報源として，名詞句で回答する質問応答システムの代表に，goo の日本語自然文検索 Web Answers[1]がある．名詞句で回答するシステムにおいては，以下のような欠点がある．

- 回答のみを示されても，ユーザは回答の正誤を判断することは難しい．
- 回答が複数存在する場合に弱く，回答を一つに絞り込もうとしてしまう．
- 「〜とは何ですか？」のように，文章で回答すべき質問は対象外となっている．

Web Answers では，回答と共に自信度と，関連する Web ページへのリンクを提示することで，上記の問題に対処しているが，名詞句ではなく，文章を回答とすることで，これらの欠点は改善される．さらに，質問応答システムの回答としては，適切な単語が示されるよりも，まとまった文章が示される方が，ユーザも回答として好むという報告もされている [4].

3　Web 文章の抜粋を回答とする質問応答システム

本研究では，文章で回答する質問応答システムを提案する．しかし，システムが自動で文章を作成することは極めて難しい．そこで，Web 上に存在する文書の膨大性から，質問の回答となる適切な文章が存在すると仮定し，Web 上の文書から適切な回答文を抽出する手法を提案する．

本システムは，質問文を入力とし，回答を含む文章を返すシステムである．質問文から，回答の種類を決定し，回答候補語を絞り込むためのルール (回答ルール) を定める．また，名詞句と動詞を抽出してキーワード群とし，キーワード群と回答の種類から検索式を作成する．キーワードと一定の範囲内で互いに共起する，回答ルールに適する語を回答候補語とする．回答候補

[1]goo ラボ, “ 日本語自然文検索, ”http://labs.nttrd.com/

語の内，多くのキーワード周辺に存在する単語であり，広く使われている回答候補語ならば，その回答候補語は解の可能性が高いとみなす．回答は，解の可能性の高い回答候補語を多く含む文章とする．本システムで回答可能な質問形式は，Where，When，Who，What 型である．

本システムの構成を図 1 に示す．

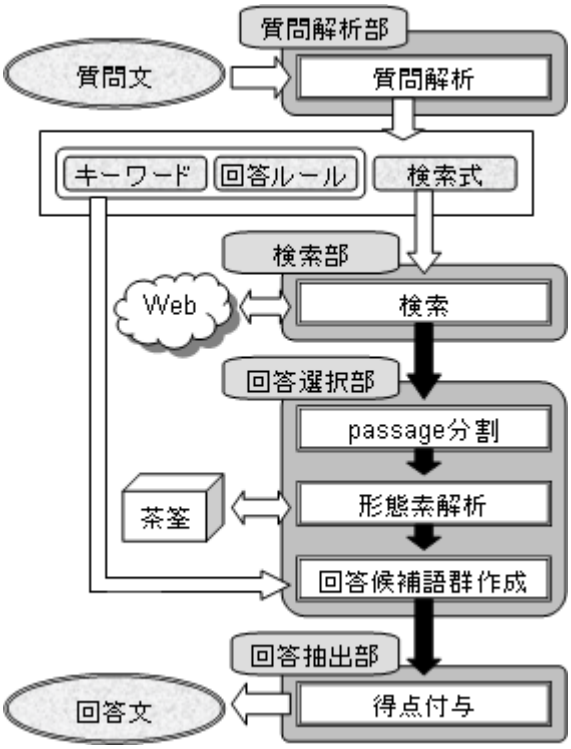

図 1: 本システムの構成

3.1 質問解析部

質問文から，検索エンジンで文書を検索するための検索式，回答抽出部において回答候補語を決定するためのキーワードと回答ルールを定める．

3.1.1 キーワードと検索式の抽出

質問文から名詞句と動詞の基本形を抽出しキーワード群とする．キーワード群から名詞句を抽出して検索式とする．

3.1.2 回答ルールの選択

質問文を Where，When，Who，What の 4W に類別し，4 通りの回答ルールを与える．

Where 質問文に「どこ」が存在する場合，回答候補語の品詞を，名詞 (地域) に限定する．

When 質問文に「いつ」が存在する場合，回答候補語を，年，月，日を単位に持つ序数詞とする．
検索式に年，月，日を加える．

Who 質問文に「誰」が存在する場合，回答候補語の品詞を，名詞 (人名) に限定する．

What 質問文に「何」が存在する場合，回答候補語は名詞，動詞，形容詞とする．
検索式の形を”〜とは”とする．

3.2 検索部

検索式を検索エンジンの Google[2]にかけ，Web から文書を取得する．Google では PageRank により期待の大きい文書から列挙されるため，上位一定数の URL を取得する．

3.3 回答抽出部

Web 文書群から適切な文章を抽出するため，Web 文書群を意味の通る小さなブロック (passage) に分割する．また，キーワードと回答ルールから，回答候補語を抽出し，回答候補語群を作成する．

3.3.1 passage 分割処理

意味的なまとまりを単位として文章を分割することは困難である．そこで，HTML のレイアウトを利用することが有用である [5]．Web 文章はタグコードにより意味的にまとまったブロックに分割されていると考え，タグコードを分割箇所として文書を passage に分割する．これにより，文書を高度に解析をせずとも，Web 文書を passage に分割することが可能である．分割箇所とするタグコートを表 1 に示す．この分割箇所を設けても文章が長いものとなる場合があるため，文字数の基準を設け，一定の文字数を超えた passage は，句読点で passage に分割する．

表 1: passage 分割箇所とするタグコード一覧

タグ	意味	タグ	意味
html	HTML 文書	ol	順序リスト
hr	横罫線	dir	リスト
br	改行	menu	メニューリスト
div	ブロック	ul	順序無しリスト
p	パラグラフ	dl	定義リスト
h	見出し	table	表作成
blockquote	引用文	tt	等幅フォント
全角スペース	段落		

[2]Google http://www.google.com/

3.3.2 形態素解析

回答候補語は，回答ルールで定めた品詞を手がかりとして抽出されるため，全ての単語の品詞を取得する必要がある．そのため，passage 群を茶筌[3]で形態素解析し，単語の基本形と品詞を取得する．

3.3.3 回答候補語群の作成

キーワードと一定の範囲内で互いに共起する，回答ルールに合う単語を，回答候補語とする．全ての回答候補語を取得し，回答候補語群とする．回答候補語は，キーワードと一定の範囲内で共起する単語で，回答ルールに合致する単語とする．

3.4 回答選択部

passage に得点をつけ，回答の可能性の高い passage から順に回答としてユーザに示す．

3.4.1 passage 選択処理

passage の中から，回答として適切な passage を選択する処理を行う．回答候補語の一定の範囲内に多くのキーワードがあり，その回答候補語が多くの passage で使われているならば，その回答候補語は解の可能性が高いとする．このような回答候補語に高得点が付くように回答候補語に得点を与え，passage に存在する回答候補語の総得点により passage に得点を与え，最高得点の passage が回答として適切な passage と判断する．

3.4.2 passage の得点計算

ある回答候補語 w と一定の範囲内で互いに共起したキーワードの数を k とする．

同一の回答候補語をまとめ, W とする．全ての $w \in W$ のキーワード数の和を求め，それを W の得点 $Point_W$ とする．すなわち，

$$Point_W = \sum_{w \in W} k$$

とする．$Point_W$ は, W の要素が多いほど, $w \in W$ のキーワード数が多いほど大きい値となる．つまり，多くの文章で，多くのキーワードと一定範囲内で共起した回答候補語が回答の可能性が高くなる．

ここで，ノイズを削除するため，一定の偏差値 T を設定し, T の場合の得点 $Point_T$ が求められるので $Point_W$ から $Point_T$ を引く．これを引いた得点である $Point'_W$ は，

$$Point'_W = Point_W - Point_T$$

となる．回答候補語 $w \in W$ の得点 $Point_w$ を

$$Point_w = Point'_W$$

とすることで，回答候補語に得点を与える．

ある *passage* の得点は，その *passage* に含まれる全ての回答候補語の得点とする．この *passage* の得点 *Point_passage* は，この *passage* に含まれている全ての回答候補語を PW としたとき，

$$Point\_passage = \sum_{w \in PW} Point_w$$

とする．これにより，回答の可能性の高い回答候補語を多く含む *passage* に高得点が付く．

4 実験結果

本システムに，Where，When，Who，What 型の質問文を与え，正解率を求めた．本実験の条件を表 2 に示す．

表 2: 実験条件

項目	条件
取得した URL 数	上位 30 件
passage の最大文字数	150 文字
キーワードと回答候補語の距離	20 単語以内
ノイズ基準の偏差値	70 以下

4.1 Where,When,Who 型に対する実験

Where, When, Who 型の質問を与え，本システム正解率を求めた．一般常識の本 [6] を参考に，Where, When，Who 型の質問文を作成し，100 件与えた時の回答となる単語を含む passage が始めて現れるまでの正解率を表 3 に示す．実験の例として，質問文，上位の回答候補群，上位 3 件の回答文を表にしたものを表 4，表 5，表 6 に示す．これら 3 つの表から，回答である単語を含む文章が回答文となっていることが分かる．表 4 からは，回答が複数個存在する場合にも適していることが分かる．

表 3: Where,When,Who 型質問への正解率

回答提示順位	正解率 (%)
1 番目	54
2 番目以内	64
3 番目以内	69
10 番目以内	75
30 番目以内	78

表 4: Where 型質問への回答例

質問文	バルト三国はどこですか？
回答候補語	エストニア_9，リトアニア_6，ラトビア_-1
回答 1	リストニア＆ラトヴィア＆リトアニア
回答 2	バルト三国はリトアニア、ラトヴィア、エストニアからなる。
回答 3	植竹繁雄外務副大臣は、１２月１７日（月）から２２日（土）まで、バルト三国（ラトビア共和国、リトアニア共和国、エストニア共和国）を訪問する。

[3]茶筌，http://chasen.naist.jp/hiki/ChaSen/

表 5: When 型質問への回答例

質問文	東京オリンピックの開会式はいつですか？
回答候補語	10 月_30，1964 年_27，10 日_18，40 年_10，8 月_-4
回答 1	１９６４年１０月１０日
回答 2	中でも有名なのが、統計上晴れる確率が高いということで昭和３９年の東京オリンピックの開会式の日となった１０月１０日の「晴れの特異日」でしょう。
回答 3	ご存知の方も多いかと思いますが、体育の日は昭和３９年１０月１０日に日本では初めての開催となる東京オリンピックの開会式を記念して制定された日（現在体育の日は、ハッピーマンデー制度により１０月の第２月曜日に）。

表 6: Who 型質問への回答例

質問文	初代アメリカ大統領は誰ですか？
回答候補語	ジョージ_17，リンカーン_-6，退助_-6，板垣_-6
回答 1	同年５月には第２回大陸会議が開かれ、６月１５日に（後の初代アメリカ大統領）ジョージ・ワシントン（４３歳。
回答 2	アメリカン大学は初代アメリカ大統領、ジョージワシントンによって発案された『アメリカの首都に偉大な大学設立』という考えに基づいて、１８９３年の議会の裁決により特許状を与えられた私立の総合教養大学で、全米から集まったおよそ６，０００人の学部生と５，０００人の院生が所属しています。
回答 3	私たち一家の住んでいるマウントバーノンエリアは初代アメリカ大統領のジョージワシントン邸（マウントバーノン）があります。

4.2 What に対する実験

What 型の質問文を与え，本システムの正解率を求めた．一般常識の本 [6] を参考に，文章が回答となる What 型の質問文を作成し，100 件与えた時の，回答となる単語を含む passage が始めて現れるまでの正解率を表 7 に示す． 実験の例として，質問文，上位の抽出

表 7: What 型質問への正解率

回答提示順位	正解率 (%)
1 件目	49
2 件目以内	66
3 件目以内	68
10 件目以内	79
30 件目以内	80

単語群，上位 3 件の回答文を表にしたものを表 8 に示す．回答 1 や回答 3 は，一般的な BSE に対する回答となっている．

表 8: What 型質問への回答例

質問文	BSE とは何ですか？
抽出単語	牛_15，する_12，脳症_10，状_10，海綿_10，病気_9，病_6，ない_6，れる_5，性_4，こと_3，よう_3
回答 1	BSE とは、どのような病気ですか？BSE ＝牛海綿状脳症（BSE:BovineSpongiformEncephalopathies）は、牛の脳組織にスポンジ状の変化を起こし、起立不能等の症状を示す遅発性かつ悪性の中枢神経の疾病です。
回答 2	なお、めん羊、山羊、ミンク等でも類似の海綿状脳症が知られていますが、これらは BSE とは異なる病気とされています。
回答 3	BSE ＝牛海綿状脳症（いわゆる狂牛病）は、牛の中枢神経が侵される病気であり、発病すると不安定な動作や運動失調が見られることが良く知られている。牛海綿状脳症は「脳がスポンジ状になる神経疾患」という意味であり、BSE に感染した牛の脳を顕微鏡で見るとスポンジのような穴が見られることから名づけられた。BSE が人に感染した場合も同じような症状（変異型クロイツフェルトヤコブ病）が見られる。

5 むすび

本研究では，Web 情報を質問応答システムの情報源としてとらえ，Web 文書群を passage に分割し，passage を回答とすることで，文章での回答を可能とした．回答が文章であるため，回答が複数ある場合や文章で答えるべき質問文にも対応できる．本研究では，上位 3 件以内に回答文が返される正解率は，Where，When，Who 型の質問では 69%，What 型の質問に対しては 68%であった．

参考文献

[1] Dragomir Radev, Weiguo Fan, Hong Qi, Harris Wu, Amardeep Grewal,“ Probabilistic Question Answering on the Web，”Proceedings of the 11th International World Wide Web Conference (2002).

[2] Stefanie Tellex, Boris Katz, Jimmy Lin, Aaron Fernandes, and Gregory Marton,“ Quantitative Evaluation of Passage Retrieval Algorithms for Question Answering, ”Proceeding of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaton Retrieval, p.41–47 (2003) .

[3] 藤井敦,“ 百科事典としての WWW，”人工知能学会誌，第 19 巻 3 号，pp.296–301 (2004).

[4] Jimmy Lin, Dennis Quan, Vineet Sinha, Karun Bakshi, David Huynh, Boris Katz, and David R. Karger“ What Makes a Good Answer? The Role of Context in Question Answering，”Proceedings of the 9th IFIP TC13 International Conference on Human-Computer Interaction(2003).

[5] 藤井敦，石川徹也,“ World Wide Web を用いた事典知識情報の抽出と組織化，”電子情報通信学会論文誌 D-II，Vol.J85-D-II No.2，pp.300-307 (2002).

[6] 短期集中　一般常識 (大学生版)，高橋書店 (2004).