Web を情報源とする日本語質問応答システムに関する研究

北斗　修哉[†]　　村田　真樹[‡]　　馬　青[†]

[†]龍谷大学大学院理工学研究科

[‡]情報通信研究機構けいはんな情報通信融合研究センター

１．はじめに

質問応答システムの研究は、新聞記事を対象とした評価プロジェクトである TREC の QA Track や NTCIR ワークショップの QAC タスクに代表されるように、これまで新聞記事を対象としたものが多くなされてきたが、我々は、さらに実用的なものとして利用できるように、Web を情報源にシステム開発を行った。情報源に Web を用いる利点として、情報量が膨大であることや、タイムリーな質問に対しても、Web 上で刻々と更新される最新の情報を情報源として活用できるなどが挙げられ、実用的な質問応答システムを研究開発する上では最適だと考えられる。Web を情報源とする質問応答については、英語を対象とした Kwok ら[1]の研究や NTT の goo サーチエンジンの検索結果を用いた日本語自然文検索実験[2]がある。本研究では、Google や Yahoo といった検索エンジンを利用し Web を情報源とする質問応答システムを、各手法を織り交ぜて開発を行った。また、評価に関して、質問応答システムの評価型ワークショップである NTCIR3 の QAC1 の Task１を用いて評価実験を行った。本稿は開発システムの詳細、評価結果、そして得られた知見について述べる。

２．システム構成

本システムの概観は図１に示すとおりで、システムの構成は、一般的な質問応答システムのそれと同様、質問解析、文書検索、回答選択の 3 つのモジュールからなる。

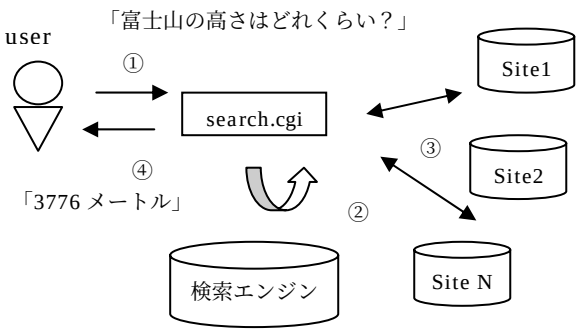

図 1　システムの概要

2.1 質問解析

質問文に含まれる「誰」や「どこ」などの疑問詞を手がかりに、人名、地名、組織名、人工名、数字の５種類の回答タイプの分類を行った。５種類の回答タイプの下位にあたる回答下位タイプの分類を、質問文の疑問詞の直前にあたる名詞と分類語彙表[3]を使って構築した。まず、疑問詞の直前にあたる名詞を取り出し、分類語彙表からその名詞の分類番号を取得する。そして、あらかじめ作成した分類語彙表の分類番号範囲とそれに対応した回答下位タイプからなる、回答下位タイプパターン辞書を使って、取得した分類番号が分類番号範囲に含まれれば、その回答下位タイプが得られる。得られた回答下位タイプは、上位にあたる５種類の回答タイプを使って取得した回答候補と比較し、一致した場合にその回答候補のスコアを上げるように、用いる。回答下位タイプは、時間、体積、距離など 20 種類程度用意した。

また、入力された質問文に対し形態素解析[4]を行い、付属語（助詞、助動詞の類）や疑問詞部分を除く表現を抽出し、これらを検索エンジンにかける検索語と、検索で得られた Web データへの回答範囲の限定や回答抽出の処理に用いる重要語とした。ただし、連語処理を行うか否かなどで得られる検索語と重要語は異なる場合がある。

2.2 文書検索

Web ページの取得には、Google Web APIs[5]と Yahoo! 検索 Web サービス[6]を使って、検索語に対する上位N件のサイトの取得を行った。取得した Web ページには『< >』で囲まれたタグと呼ばれる命令が記述されており、本研究ではタグを除去しテキスト化した。テキスト化された Web ページに対して、質問文に対する回答が書かれていそうな部分の抽出を行うため文書検索を行う。文書検索では２段階の処理を行っている。まず、各サイトのテキスト化されたデータを一行ずつ読み込み、重要語が含まれていれば、その行を含む上下ｎ行を連結し、１つのブロックとする。

このように得られたブロックはサイトごとに取っておく。次に得られた全ブロックに対して式（1）を使って上位 N 件のブロックを抽出し回答抽出対象とした。

$$score(Bi) = \sum_{t \in Bi} idf(t) \quad (1)$$

ここで Bi と t はそれぞれブロックと重要語を表す。また $idf(t)$ の値はあらかじめ数年分の新聞記事を使って求めておいた重用語の重みとして使うものである。

2.3 回答選択

回答抽出対象の各ブロックから回答タイプに合う固有表現の抽出を行い、抽出された固有表現を回答候補とした。固有表現抽出は、先行研究[7]にならって、固有表現キーと形態素解析の品詞情報を使って行った。各サイトのブロックごとの回答候補に対して式（2）を使ってスコア付けを行った。

$$score(c) = \sum_{t \in q} \frac{1}{dist(c,t)} \times idf(t) \quad (2)$$

ここで c と t はそれぞれ回答候補と重要語を表し、$dist(c,t)$ は回答候補と重要語との距離を表す。$idf(t)$ とは重要語の重みである。

各サイトの回答候補に対して、同じ回答候補があれば、加算処理を行う。同じサイト内の重複する回答候補には、加算処理を行わず、最もスコアの高いものをそのサイトからの回答候補としている。加算方法には単純加算による方法と逓減加点[8]による方法を用いている。単純加算とは、同じ回答候補のスコアを加算して、その和を回答候補の総スコアとするものである。逓減加点法とは、先行研究で行われている、得点を減らしながら式（3）を用いた加算処理を行うものである。

$$total\_score(c) = \sum_{i=2}^{n} k^{i-1} score(i) \quad (3)$$

すなわち、重複する回答候補 c に対して、スコアを降順にソートし、i 番目の候補の得点には k^{i-1} の値を乗じてから加算する方法である。重複する回答候補の処理は単純加算と逓減加点のいずれかを使用して行っている。また、k の値は実験で定めている。

また、係り受けを使った加点処理も行っている。例えば、「日本の平均寿命はどれくらいですか？」といった質問文に対して『日本の平均寿命は去年より０．２８歳延びた、７８．６４歳です。‥‥』となるブロックが存在した場合、重要語と回答候補との単語間距離だけでは、『７８．６４歳』より『０．２８歳』の方が重要語との距離が近く、回答として上位にあがってしまう。そこで、係り受け関係を使った回答選択を行った。具体的には、質問文の疑問詞の直前にあたる名詞を使って、各ブロックに対して係り受け解析[9]を行い、直前の名詞を含む文節が掛かる文節からの固有表現抽出と直前の名詞を含む文節へ掛かる文節からの固有表現抽出を行う。このように係り受け関係にある文節のみを使って回答抽出を行った。

さらに、回答表現の意味制約に基づく加点処理も行っている。例えば、「富士山の高さは何メートルですか？」といった質問のように、回答候補が「○○メートル」「○○m」のような「何＋名詞」となる質問に対しては、語尾に名詞を持つ回答候補に対して、加点を行っている。

以上の手法を織り交ぜ Web を情報源とした質問応答システムを構築した（図 2）。

図 2　検索結果画面の例

３．実験

実験には NTCIR3 の QAC1 の Task1 を利用した。具体的には NTCIR の評価データを使って評価するとともに、NTCIR の質問を使って情報源を Web に変えて評価を行った。検索エンジンからの Web ページの取得に関しては、検索結果の上位 20 サイトを使い、上位 10 ブロックのデータを使用し回答の出力を行った。

3.1 評価方法

MRR と 5 次正解率を用いて評価を行った。また、Web での評価に際して NTCIR の評価データは本来新聞記事を対象としたもので、過去の新聞記事から構成された正解データと Web ページから得られた最新な情報に表記や回答

表 1　タイプ別の Google、Yahoo、新聞記事での精度結果

		Google		Yahoo		新聞記事	
質問タイプ	質問数	MRR	5 次正解率	MRR	5 次正解率	MRR	5 次正解率
人名	43	0.359	46.5%	0.239	34.9%	0.418	46.5%
地名	29	0.508	58.6%	0.439	51.7%	0.436	51.7%
組織名	19	0.434	47.4%	0.405	47.4%	0.399	52.6%
人工名	62	0.404	51.6%	0.450	51.6%	0.341	45.2%
数字	47	0.374	46.8%	0.373	42.6%	0.372	46.8%
total	200	0.405	50.0%	0.381	45.5%	0.384	47.5%

が異なる場合もあるので修正を行った。新聞記事では全角文字で統一されているが Web ページでの表記は様々なので、例えば、正解データ「５２．６０%」に対して「52.6%」や「５２.６パーセント」などの半角文字や単位表現の言い換えなどを正解データとした。また、質問文と Web ページ内の文脈を吟味し、例えば「盲導犬は日本全国にどのくらいいますか。」といった質問に対して、正解データは「約８００頭」や「８００頭余り」となっているが、Web ページ内の記述より、「現在、街で活躍する盲導犬の数は、およそ 900 頭で、日本の盲導犬の頭数は圧倒的に少ない」の文脈から「およそ 900 頭」を正解データとして修正を行った。

3.2　実験結果と考察

Google、Yahoo、新聞記事のそれぞれを使った質問応答の精度を表１に示す。Web を利用した場合の問題点として、「表記のゆれ」が多く、例えば「333m」「３３３m」「三百三十三メートル」「三三三メートル」といったこれらの回答候補を重複回答として加算することができず、とくに「人名」や「数字」での「表記のゆれ」が多く出現した。また、質問文が長くそのため検索語が多く抽出された場合、検索エンジンで検索を行うと、ヒット数が 0 となるケースが Google で９問、Yahoo で１３問存在し、これらの質問に対しては、回答を提示することができなかった。また、Google で 5 位以内に回答があり、Yahoo では 5 位以内に回答がないような質問は 21 問あり、逆に、Yahoo では 5 位以内に回答があり、Google で 5 位以内に回答がないような質問は 13 問あった。これら Google と Yahoo での5位以内の正解数を合わせると113問となり何らかの手法によって、併用や組み合わせることで、回答のさらなる精度向上に繋がると考えられる。

次に各システム別の精度を示す。我々は 15 種類のシステムを用意して比較実験を行った。それぞれのシステムは 6 種類の設定（単純加算、逓減加点、係り受け、回答下位タイプ、IDF、分類語彙表）の組み合わせからなっている（表 2）。使用した場合のオプションを Y または定数、使用しない場合は N で表している。単純加算、逓減加点とは重複する回答候補の加算方法で、逓減加点の列に書かれた定数は式（3）の定数kの値で、係り受けとは、係り受けを考慮した加点法である。また、IDF とは式（1）（2）で用いたもので、使用しない場合は単語の重みをすべて１とした。また、分類語彙表とは、重要語に、重要語と同じカテゴリーで隣接する単語を分類語彙表から取得し追加したものである。システム別の精度を求めたグラフを図 3 に示す。グラフの X 軸には、15 種類のシステムを、Y 軸には

表 2　６種類の設定からなる各システム

	sys1	sys2	sys3	sys4	sys5	sys6	sys7	sys8	sys9	sys10	sys11	sys12	sys13	sys14	sys15
単純加算	Y	Y	Y	N	N	N	N	N	N	N	N	N	Y	Y	Y
逓減加点	N	N	N	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	N	N	N
係り受け	N	N	N	N	N	N	N	N	N	N	N	N	N	Y	Y
回答タイプ	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
IDF	Y	N	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N
分類語彙表	N	N	N	N	N	N	N	N	N	N	N	N	Y	N	N

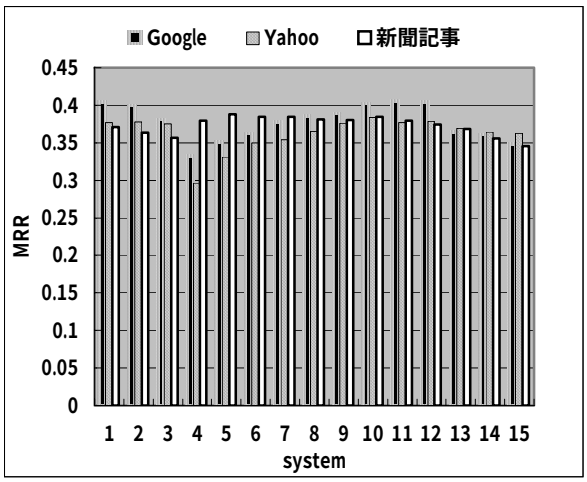

図 3 Google、Yahoo、新聞記事での各システムの精度

MRR 値を表している。

各システムを比較して見ていくと、まず、単純加算と逓減加点の加算方法について、新聞記事を利用した場合は逓減加点での結果が単純加算での結果よりも 0.01 程度、Web でも、0.001 程度精度がよかった。また、逓減加点法は、ブロック数を上位 10 件に絞らず、上位 100 件とブロック数を増やすと、単純加算では精度を大きく下げてしまうが逓減加点では、図には示していないが、新聞記事と Yahoo でそれぞれ、system6 で MRR が 0.394 と system10 で MRR が 0.387 と上位 10 件のブロックを使用した場合のもっとも良い結果よりさらに精度が上がった。しかし、すべてのブロック数を増やせばそれだけ解析時間が掛かってしまうデメリットもあり、用途によっての使い分けが考えられる。また、逓減加点で使用した係数には、新聞記事では 0.2 や 0.3 の場合の結果が一番よく、新聞記事を対象としている先行研究と同様の結果となった。一方 Web での場合は、逓減加点で使用した係数には、0.7 や 0.8 の場合の結果が一番よいことが分かった。次に、回答タイプを細分化した場合としない場合との比較では、新聞記事、Web どちらにおいても system2 と system3 との比較により、回答タイプを細分化した system2 の方が、精度が高くなった。また、重要語の重みについては、system1 と system2 を比較した場合になるが、新聞記事では IDF を使った方が結果がよく、Web では Yahoo を使った場合、僅かながら、IDF を使った方が精度を下げる結果となった。Web を用いた場合の、重要語の重み付けについては、さらに検討が必要だと思われる。また、係り受けについては、係り受けを使用しないものと比べて使用する場合は、精度を下げてしまった。原因を調べたところ、質問文の疑問詞の直前にあたる名詞を使った係り受け関係からでは、正解の回答を含む文節が係り受け関係になく、回答を取得できないケースなどがあり改善が必要となった。

４．おわりに

Web を情報源とする質問応答システムの開発を行った。Web を情報源とした場合、質問文から抽出したすべての検索語を使って検索をかけると、ヒットしない場合があることや、回答表現として、「表記のゆれ」が多く存在し、重複回答を加算する処理で、回答候補を同一として扱えないため得点が加算できない場合など確認された。今後の課題として各モジュールでの精度の向上を図るとともに、上記で述べた課題に取り組んで行く予定である。

参考文献

[1] Kwok, C. C. T., Etzioni, O. and Weld, D. S.: Scaling Question Answering to the Web, WWW10 (2001)

[2] goo ラボ：日本語自然文検索実験、http://labs.nttrd.com/

[3] 国立国語研究所：分類語彙表、大日本図書、1964 年 3 月

[4] 松本 裕治, 今一 修, 山下 達雄, 北内 啓, 今村 友明：“日本語形態素解析システム 茶筌 version 2.2.9 使用説明書 ”、奈良先端科学技術大学院大学松本研究室, 2001

[5] Google Web APIs：http://www.google.com/apis/

[6] Yahoo! デベロッパーネットワーク：http://developer.yahoo.co.jp/

[7] 渡辺一郎　桝井文人 福本淳一：固有表現抽出ツールＮＥｘＴの精緻化とユーザビリティの向上、言語処理学会第 10 回年次大会、2004 年 3 月

[8] 村田 真樹 内山 将夫 井佐原 均：質問応答システムにおける逓減加点法に基づく複数記事情報の利用、自然言語処理研究会、2004 年 3 月

[9] 日本語係り受け解析器 CaboCha 「南瓜」：奈良先端科学技術大学院大学松本研究室