日本語自然文検索システム Web Answers

永田昌明　　齋藤邦子　　松尾義博

日本電信電話株式会社 NTT サイバースペース研究所
{nagata.masaaki,saito.kuniko,matsuo.yoshihiro}@lab.ntt.co.jp

1　はじめに

近年、Mulder[3], AnswerBus[6], NSIR[4], AskMSR[1] など、Web を知識源とする質問応答システムが盛んに研究されている。また、AskJeeves, MSN Search, Google などの商用インターネット検索エンジンは、簡単な事実を問う自然言語の質問文に対して直接回答を返す機能を提供し始めた。

本稿では、Web 検索エンジンの自然文検索インタフェースとして設計された実時間 Web 質問応答システム “Web Answers” について報告する。本システムは自然言語で入力 された検索要求に対して適合する Web 文書を提示するだけでなく、名称・数量・ 定義・評判を問う質問に対して、可能ならば検索された文書から回答を抽出して表示する。本システムは質問パターンと回答パターンの対からなる語彙意味規則により動作を記述し、大規模なシソーラスと高速な固有表現抽出器を利用することにより広範な適用領域と実時間応答を達成した。

図 1 に Web Answers の構成を示す。以下では、まず本システムが対象とする質問応答の範囲およびその実現方法について説明し、次に公開実験で収集した質問文を用いた本システムの精度評価の結果について報告する[1]。

2　質問応答の対象範囲

質問の種類による質問応答の難しさの違いを議論するために、回答の客観性と回答の長さを分類軸として、質問を分類したものを図 2 に示す。一般に質問応答は、名称や数量のように正解が客観的に一つに決まり、かつ回答が短い単語列である場合が最も易しい。逆に評判のように正解が存在しない場合や、方法のように回答が長い説明文になる場合ほど難しい。

Web Answers では名称・数量に加えて、簡潔な句または文を回答とできる場合が多い定義・評判を処理対象とする。これは、名称・数量を問う質問 (factoid) だけを処理対象として公開実験を始めたところ、「自分が回答を知っている質問を入力してシステムの回答を楽しむ」という使い方が多かったので、non-factoid な質問応答を実現して違うユーザ層へ訴求しようという狙いもある。

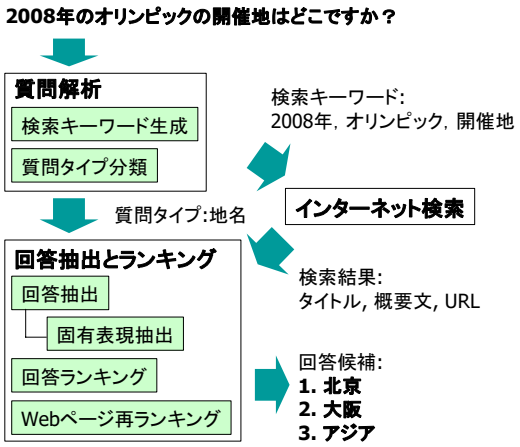

図 1: Web Answers の構成

3　質問解析

3.1　検索キーワード の生成

質問解析部は検索キーワード の生成と質問タイプの分類を行なう。まず形態素解析器 [5] により質問文の単語分割と品詞付与を行ない、ストップワードを取り除いた残りの単語をインターネット検索エンジン goo (goo.ne.jp) へ送り、タイトル・概要文 (snippet)・URL からなる上位 10 個の検索結果を得る。例えば図 1 において「2008 年のオリンピックの開催地はどこですか?」という質問文から「2008 年」「オリンピック」「開催地」が検索キーワードとして抽出される。

質問文が長くなるほど検索キーワード の数が増えるの

[1] 本システムは「日本語自然文検索実験」として goo ラボ (labs.goo.ne.jp) で 2004.2.5 から 2005.5.9 まで一般公開された。

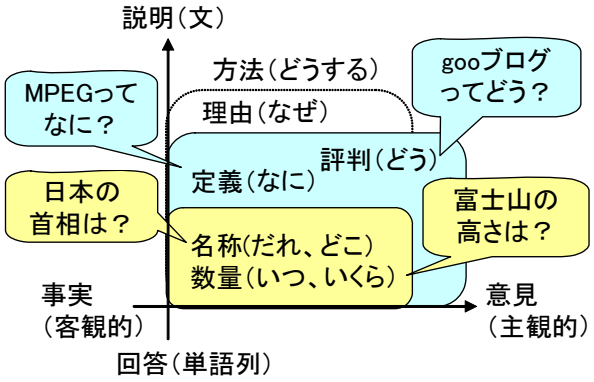

図 2: Web Answers が対象とする質問応答の範囲

```
# Q どこの{組織 (362)}
#   何という{組織 (362)}
#   {組織 (362)}は何 (どこ) ですか
# A <ORG> 1 単語以上
#   <LOC> 店舗等 (430)
#   名詞:nt, kata
qpat=<h:(?:何処|どこ)><h:の><b:362>
qpat=<h:何><h:と><h:い><h:う><b:362>
qpat=<b:362><h:は><h:(?:何|何処|どこ)>
apat=(<n:B-ORG>(?:<n:I-ORG>)*)
apat=(<n:B-LOC/b:430>)|
      <n:B-LOC>(?:<n:I-LOC>)*(?:<n:I-LOC/b:430>))
apat=((?:<p:(?:名詞:nt|kata)>)+)
answer=%1%
```

図 3: 質問回答規則の例

で、逆文書頻度 (IDF) を重要度スコアとして検索キーワードを最大 5 個までに絞り込む。なお IDF は約 5ヶ月分の新聞記事から求めた。

3.2 質問タイプの分類

次に質問解析部は質問タイプを決定する。質問タイプは入力された質問文が求めている回答に基づく質問文の分類である。質問タイプは約 50 個の人手で作成された質問回答抽出規則により決定する。質問回答抽出規則は、質問パターンと回答パターンの組から構成される。各パターンは単語の素性を指定可能な perl 風の正規表現であり、もし入力された質問文が質問パターンと照合したら、対応する回答パターンが回答抽出に使用される。

図 3 に質問回答規則の例を示す。qpat と apat はそれぞれ質問パターンと回答パターンを表す。answer は回答である。'%1%' は正規表現照合の後方参照を表し、回答パターンにおいて'(' と')' で囲まれた部分に対応する。

単語は'<' と'>' で囲み、名前と値の対を':' で区切って素性を指定する。素性名としては h, p, n, b などが指定でき、それぞれ表記・品詞・固有表現タグ・意味カテゴリを表す。一つの単語に対して 2 つ以上の素性を指定する場合は'/' で素性を区切る。

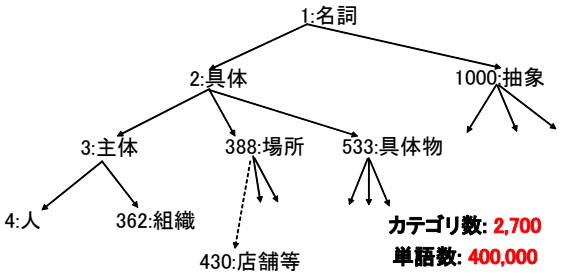

図 4: 日本語語彙大系の意味カテゴリ

意味カテゴリは、国内最大の日本語シソーラスの一つである日本語語彙大系 [2] の意味カテゴリ番号で指定する。日本語語彙大系の上位階層と意味カテゴリ番号を図 4 に示す。日本語語彙大系は約 40 万語の語彙が約 2700 のカテゴリに分類されている。

例えば, 素性として b:362 を指定すると、「組織」およびその下位カテゴリに登録されているすべての単語がこの正規表現と照合する。従って、図 3 の最初の qpat は、「どこの会社」や「何処のチーム」など、疑問文の主辞名詞の意味カテゴリが「組織」またはその下位カテゴリである質問文と照合する。

4 回答抽出とランキング

4.1 回答抽出

回答抽出部は、まず隠れマルコフモデルに基づく固有表現抽出器 NameLister を用いて、検索結果の上位 10 件の概要文 (snippet) の単語分割・品詞付与・固有表現タグ付与を行ない [5]、次に質問文と照合した質問パターンに対応する回答パターンを用いて概要文から回答候補を抽出する。

図 5 に NameLister の出力例を示す。第 1,2,3 列はそれぞれ表記・品詞・固有表現タグを表す。「巨人の松井秀喜外野手は 11 月 1 日に東京都内で記者会見し...」という文において、「巨人」「松井秀喜」「11 月 1 日」「東京都」がそれぞれ ORG(組織名),PSN(人名),DAT(日付),LOC(場所) として抽出される。一つの固有表現が 2 つ以上の単語から構成される場合や、同じ固有表現クラスの固有表現が 2 つ以上連続して出現する場合に対応するため、各単語の固有表現タグは、'B-' または'I-' という接頭辞と

```
巨人        名詞:nt      B-ORG
の          格助詞       NIL
松井        名詞:nr      B-PSN
秀喜        名詞:nr      I-PSN
外野手      名詞         NIL
は          連用助詞     NIL
１１月      名詞         B-DAT
１日        名詞         I-DAT
に          格助詞       NIL
東京        名詞:ns      B-LOC
都          名詞接尾辞   I-LOC
内          名詞接尾辞   NIL
で          格助詞       NIL
記者会見    名詞         NIL
し          動詞接尾辞   NIL
```

図 5: NameLister の出力の例

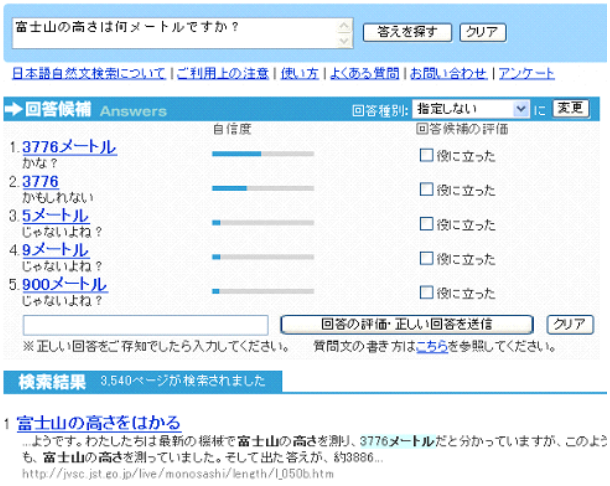

図 6: Web Answers のユーザインタフェース

固有表現クラスの組合せになっている。ここで'B' は固有表現の先頭の単語を表し、'I' は固有表現の先頭以外の単語を表す。図 3 では最初の apat が固有表現タグの使用例になっており、これは一つの単語 (B-ORG だけ) または二つ以上の単語 (B-ORG の後に I-ORG を繰り返す) から構成される組織名を表す。

定義を問う質問に対しては、名詞述語文や連体節 (句) に含まれる定義表現 (X とは〜である, 〜した X, 〜の X, など) を回答候補として抽出する。評価を問う質問に対しては、予め評価表現辞書を用意し、評価表現を中心とする構文パターンを回答候補として抽出する。

4.2 回答ランキング

抽出された回答候補は、概要文における出現位置、出現頻度、回答候補と検索キーワードの距離などに基づいて順位付けを行なう。上位の検索結果において検索キーワードの近くに数多く出現した回答候補ほど正解である可能性が高いと仮定し、以下のような回答候補スコア Sa を定義する。

$$Sa = \sum_{i=1}^{10} \frac{w_i}{Ns_i} \sum_{j=1}^{Ns_i} \frac{1}{Nqw} \sum_{k=1}^{Nqw} \exp(-d_{jk}^2/C) \qquad (1)$$

ここで i は検索エンジンによる文書の順位であり、Ns_i は i 番目の概要文中の文数である。Nqw は検索キーワードの数であり、d_{jk} は文 j における検索キーワード k と回答候補の最短距離を表す。w_i は i 番目の概要文の重みであり ($\sum w_i = 1$)、C は距離に対する重みである。w_i および C は実験により決定する。

定義や評判を問う質問の場合、同じ表現が概要文に複数回出現する可能性は低いので、前記の回答候補スコア Sa のほかに、文脈に基づく「定義らしさ」「評判らしさ」を考慮する。「定義らしさ」および「評判らしさ」は、百科事典・掲示板・ブログなどから収集した定義および評価を含む典型的な文を学習データとするテキスト分類器により判定する。

4.3 Web ページの再ランキング

本システムでは、質問に対する回答を含む可能性の大きさに基づいて Web ページを独自に順位付けする。概要文と質問文が同じ表現を多く含み、かつ、概要文が尤もらしい回答候補を多く含むほど、その文書が回答を含む可能性が高いと仮定し、以下のような Web ページスコア Sd を定義する。

$$Sd = w_1 \sum_{i=1}^{N_1} tf_i + w_2 \sum_{j=1}^{N_2} tf_j + w_3 \sum_{k=1}^{N_3} tf_k + w_a \sum_{l=1}^{N_a} Sa_l \qquad (2)$$

ここで $N_n (n = 1, 2, 3)$ は質問文中の n-gram の異なり数である。tf_n は概要文中の n-gram の頻度を表し、w_n は n-gram の重みである。N_a は概要文中の回答候補の数であり、Sa_l は式 (1) で定義される回答候補スコアである。重み w_n と w_a は実験により決定する。

名称	46.3	人名 (22.0), 地名 (6.1), 組織名 (4.7), サイト (3.0), 固有物名 (4.4), その他 (5.9)
数量	9.3	日付 (3.6), 時刻 (0.1), 時間 (0.3), 期間 (0.2), 金額 (1.5), その他 (3.6)
説明	27.4	原因 (0.5), 原理 (0.2), 理由 (0.7), 方法 (4.3), 意味 (7.7), 正体 (5.9), 評判 (1.4), 連想 (0.5), その他 (6.2)
真偽	4.7	
その他	11.3	

表 1: 要求される回答に基づく質問文の分類 (単位:%)

5 ユーザインタフェース

図 6 に「富士山の高さは何メートルですか?」という質問文に対する Web Answers の出力画面を示す。ユーザが自然言語による検索要求を画面上段のテキストボックスへ入力して「答を探す」ボタンを押すと、上位 5 個の回答候補が中段に表示され、再ランキングされた上位 10 件の検索結果が下段に表示される。

式 (1) で求めた各回答候補の確信度は、水平方向の棒グラフ、および、回答候補の直下の自然言語により表示される。ここでは確信度 Sa の値に応じて以下のような表現を選んでいる[2]。

$1.0 \geq Sa \geq 0.8$ ちがいない
$0.8 > Sa \geq 0.6$ だよね?
$0.6 > Sa \geq 0.4$ かな?)
$0.4 > Sa \geq 0.2$ かもしれない
$0.2 > Sa \geq 0.05$ じゃないよね?
$0.05 > Sa \geq 0.0$ わけないか...

またユーザは各回答候補の横にある「役に立った」というチェックボックスをチェックするか、回答候補の直下にあるテキストボックスに回答を入力して「回答の評価・正しい回答を送信」ボタンを押すことにより、システムにフィードバックを返すことができる。

6 名称と数量に関する精度評価

一般ユーザが入力する質問文は TREC や QAC などの質問応答に関する学術的コンテストのテスト文よりもはるかにバラエティに富んでいるため、Web を知識源とする質問応答の精度評価は非常に難しい。

まず適当に選んだ一日 (2004 年 2 月 12 日) に公開実験サイトに入力された全ての質問文 (32,304) を、要求される回答の種類に基づいて表 1 のように分類した。その結果、全体に占める割合が多く、かつ、回答を容易に準備できるという理由から、名称と数量を問う質問文に限定して精度評価を行なうことにした。名称と数量を問う質問から、下位分類の割合が表 1 に比例するよう選んだ 200 個のテスト文をテストセット WA040212 と呼ぶ。

[2]掲示板やブログの反響をみると、多くのユーザはこの自然言語による確信度表示に親しみを感じるようである。

	MRR1	MRR2	AnsInSnippet
WA040212	0.220	0.355	0.620

表 2: 名称と数量に関する質問応答の精度

表 2 に本システムの WA040212 に対する平均逆順位 (MRR, Mean Reciprocal Rank) を示す。MRR1(=0.220) は全ての質問文 (200 個) に対する MRR であるのに対し、MRR2(=0.355) は上位 10 件の概要文に正解を含む質問文 (124 個) に限定した MRR である。AnsInSnippet(=0.620) は検索結果の上位 10 件の概要文に正解を含んでいた割合を示す。MRR1 は知識源としての Web と本システムの両方の評価であり、MRR2 は本システムの回答抽出部の評価である。MRR1 が 0.22 ということは、平均的には 4 位と 5 位の間に正解が表示されるということを意味する。取り扱う話題の広さと応答の速さを考慮すると、これは reasonable な精度だと我々は考えている。さらに詳しい誤り分析を行なったところ、誤り原因の 30%は質問解析にあり、50%は回答抽出にあることが分かった。

7 おわりに

本稿では、大規模なシソーラスと高速な固有表現抽出器を利用することにより、Web を知識源として広範な質問に実時間で回答する質問応答システムを実現できることを示した。今後の課題は、定義および評判に関する質問応答の精度を評価することである。

参考文献

[1] S. Dumais, M. Banko, E. Brill, J. Lin, and A. Ng. Web question answering: Is more always better? In *Proceedings of SIGIR-2002*, pp. 291–298, 2002.

[2] 池原, 宮崎, 白井, 横尾, 中岩, 小倉, 大山, 林 (編). 日本語語彙大系. 岩波書店, 1997.

[3] C. C. Kwok, O. Etzioni, and D. S. Weld. Scaling question answering to the web. In *Proceedings of WWW-10*, pp. 150–161, 2001.

[4] D. Radev, W. Fan, H. Qi, H. Wu, and A. Grewal. Probabilistic question answering on the web. In *Proceedings of WWW-11*, pp. 408–419, 2002.

[5] K. Saito and M. Nagata. Multi-language named-entity recognition system based on HMM. In *Proceedings of ACL2003 Workshop on MuLNER*, pp. 41–48, 2003.

[6] Z. Zheng. Answerbus question answering system. In *Proceedings of HLT-2002*, 2002.