2007－NL－182 (15)
2007／11／20

複数の Web 検索エンジンを用いた factoid 型質問応答

金井 明
横浜国立大学 大学院 環境情報学府

佐藤 充
横浜国立大学 大学院 環境情報学府

石下 円香
横浜国立大学 大学院 環境情報学府

森 辰則
横浜国立大学 大学院 環境情報研究院

E-mail: {a-kanai,mitsuru,ishioroshi,mori}@forest.eis.ynu.ac.jp

ネットワーク上には複数の異なる商用 Web 検索エンジンが存在し，広く利用されている．Web 文書を情報源とする質問応答 (QA) においては，異なる Web 検索エンジンの出力結果を組合せることによって，求解精度の向上が期待される．それは，各検索結果が同一であることがまずないために，それらを組合せることにより情報源の多様性を増すことができるためである．しかしながら，QA において異なる Web 検索エンジンを組合せることによる効果の研究は今までなされていなかった．

そこで本論文では，QA における，異なる Web 検索エンジンを組合せることによる効果について調査した．具体的には，異なる検索エンジンから得られた検索結果を組合せるタイミングに注目し，異なる 3 手法を比較検討した．1 つ目の手法は各検索エンジンの出力結果をそのまま併合し，QA エンジンで処理する従来手法である．2 つ目と 3 つ目は，我々の提案手法であり，各検索エンジンの出力結果をそれぞれ個別に QA エンジンで処理し，そこで得られた解候補を併合する方法である．評価実験によれば，質問応答処理の前に検索結果を併合する従来手法よりも，QA エンジンによる個別の質問応答処理の後に，抽出された解候補を組合せるほうが効果ががあることが示された．

Effect of combining different Web search engines on Web question-answering for factoid questions.

Akira KANAI Mitsuru SATO Madoka ISIOROSHI Tatsunori MORI
Graduate School of Environment and Information Sciences,
Yokohama National University
E-mail: {a-kanai,mitsuru,ishioroshi,mori}@forest.eis.ynu.ac.jp

We have several different commercial Web search engines that are available. It is expected that the combination of the results from different Web search engines has some impact on the accuracy of question-answering (QA) for Web documents, because the search results are not identical and the combination increase the variety of information source. However, as far as we know, there is no studies on the effect of combining different Web search engines on QA.

In this paper, we examined the effect of the combination on QA. We proposed three different methods to combine search results from different search engines in the process of QA. The first one is the method that straightforwardly merges search results from different search engines, then, feeds the unified search result into one QA engine. On the other hand, the second one and third one are methods that feed each search result from individual search engine to a QA engine separately, then merge the answer candidates. The experimental result showed that the methods that merge the answer candidates after QA is more effective than the method that merge the search results before QA.

1 はじめに

質問応答 (QA) は，情報検索 (IR) 技術ならびに情報抽出 (IE) 技術の組合せの発展形として，広く研究が行なわれている．QA システムは質問に対して関連文書を探して提示するだけではなく，直接質問の答を提示してくれる．例えば，「日本の首都はどこですか?」という質問文に対し，質問応答システムは，情報源から解候補を探し出し，「東京」という答を返すことが期待されている．

情報源の観点から述べると，初期の質問応答システムでは，新聞記事集合のような静的かつローカルな文書群を用いていたが，文書の範囲や数が限られるとともに，文書集合が固定で更新されないため，システムの回答できる質問が当時の時事に限られるという制約があった．そこで，近年では，文書が豊富で日々追加・更新がなされる Web 文書に注目し，これを情報源とした質問応答システム (Web QA) が研究されている．通常，Web 検索エンジンを質問応答用に独自に用意することは現実的でないので，

既存のWeb検索エンジンが利用される．

ここで，我々は複数の異なる検索エンジンがWeb QAに使用できるということに注目している．各検索エンジンの出力する検索結果が同一であることはまずなく，それらを組合せることにより，情報源の多様性を増す事ができる．そのために，異なるWeb検索エンジンの検索結果を組合せることにより，Web QAの精度向上に何らかの効果があると期待される．

しかしながら，QAにおいて異なるWeb検索エンジンを組合せることによる効果の研究は，我々の知る限り，今までなされていなかった．そこで本論文では，その効果について調査を行なった．具体的には，異なる検索エンジンから得られた検索結果を組合せるタイミングに注目し，異なる3手法を比較検討した．1つ目の手法は各検索エンジンの出力結果をそのまま併合し，QAエンジンで処理する従来手法である．2つ目と3つ目は，我々の提案手法であり，各検索エンジンの出力結果をそれぞれ個別にQAエンジンで処理し，そこで得られた解候補を併合する方法である．

2 関連研究

Linら[6]によると，Web文書を利用する方法には少なくとも2つの種類がある．一つは，Web文書を主情報源として扱う方法である．もう一つは，Web文書を，主情報源とする他のコーパスと組合せて使用する方法である．本論文では，前者の，Web文書を主情報源として扱う方法を考える．

Web文書を知識源として利用する一つの利点として，Web文書におけるデータの冗長性が挙げられる．自然言語の表現能力の高さは，同じ事柄を多様な表現で表すことを許している．そのため，ある質問文に対する答が，その質問文とは異なる表現の文脈に現れる可能性が高い．質問文と情報源の間に存在するこの種の齟齬は，質問応答において適切な解候補を見つけ出す際に，主要な障害の1つとなる．

しかし，Webにおけるデータの冗長性に注目すると，ある質問文に対する(正しい)解に対応する語句について，Web上では多数の異なる著者が各々異なる表現方法でその説明を行なっていることが予想される．このとき，質問文が問うているのと非常に似通った表現でその語句の説明を行なっている文書が存在する可能性が高くなると期待できる[6]．すなわち，Webにおけるデータの冗長性により，表現の多様性を確保でき，質問文と非常に似通った文脈が現れるため，質問応答処理において言い換え等の高度な言語処理を行なわずともある程度の求解精度が確保できると期待される．

表現の多様性についていえば，いくつかの研究において情報源の多様性が活用されている．例えば，最初のWeb QAシステムの1つであるSTARTの最近の版では，複数の情報源を活用している[4, 5]．Radevら[9]はWeb QAにおける確率に基づくアプローチを提案しているが，そこでは3つの主要なWeb検索エンジンを組合せて上位40文書を得ている．

これらの研究では異なる情報源から得られた文書を使用してはいるが，文書検索よりも後の段階においては，情報源の違いは考慮していない．これに対して，節4で述べる我々の手法では，異なる情報源の間でのデータの冗長性を活用しようとしている．

3 基本となる質問応答システム

本研究で使用するQAシステムは森[7]に基づく実時間QAシステムである．これは日本語のfactoid型質問に対し，日本語で答を返すシステムである．図1に示すとおり，このシステムは質問文解析(question analysis)部，外部検索エンジンへのインターフェース(interface to external search engine)，パッセージ抽出(passage extraction)部，文照合(sentential matching)部，解生成(answer generation)部，疑似投票(pseudo voting)部から構成されている．

質問文解析部は利用者から質問文を受け取り，キーワードのリストや質問文の型などの情報を抽出する．本論文では，**キーワード**を質問文中の内容語と定義する．キーワードのリストが検索質問として外部検索エンジンに入力され，関連文書群が検索される．

文照合部は，文書集合から抽出された文集合をパッセージ抽出部から受けとり，それらを処理する．ここで得られた各文を本論文では**検索文**と呼ぶ．各検索文中の各形態素が一つの解候補として扱われ，それらに対して，次節に述べる方法によりスコアが与えられる．ここで，ある形態素がある単語自身であったり，より長い複合語の一部であったりすることに注意されたい．そのため，後者の場合には，解生成部において，その解候補となる形態素を含む複合語を抽出し，それを改めて解候補として取り扱う．

3.1 解候補のスコア付け

基本となるQAシステムでは，式(1)に示す，解候補に対する複合的な照合スコアを採用している．

$$S(AC, L_i, L_q) = Sb(AC, L_i, L_q) + Sk(AC, L_i, L_q) + Sd(AC, L_i, L_q) + St(AC, L_i, L_q) \quad (1)$$

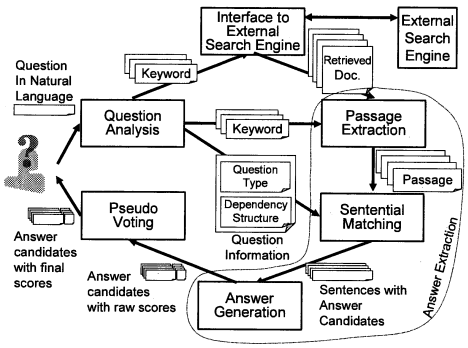

図 1: 基本となる質問応答システム

本論文ではこのスコアのことを**原スコア** (raw score) と呼ぶ．このスコアは，i 番目の検索文 L_i にある解候補 AC に対し，質問文 L_q に関する以下の 4 つの部分スコアを計算し，その線形結合を求めたものである．

1. $Sb(AC, L_i, L_q)$: 文字 2-grams の観点で計算した照合スコア
2. $Sk(AC, L_i, L_q)$: キーワードの観点で計算した照合スコア
3. $Sd(AC, L_i, L_q)$: 解候補とキーワードの間の依存構造の観点で計算した照合スコア
4. $St(AC, L_i, L_q)$: 質問文の型の観点で計算した照合スコア

$St(AC, L_i, L_q)$ の計算では，IREX-NE タスク [3] で定義された 8 種類の固有表現型を識別できる固有表現抽出器を用いた．

なお，スコア計算に纏わる計算量の削減のために，文照合部には A* に基づく探索制御が導入されている．この制御手法により，システムは最も有望な解候補のスコア計算を優先的に行ない，それ以外の解候補のスコア計算を遅延させることが可能となり，解候補の n-best 探索ができる．

3.2 探索制御の枠組における疑似投票手法

既存の多くの QA システムでは解候補に関する大域的な情報を利用している．特に冗長性は最も基本的であり，かつ重要な情報である．例えば，文書中に複数回出現する解候補に対し，そのスコアを増加させるという手法がある．これは，投票手法 (voting method) として知られる [1, 10].

一方で，我々の QA システムのように解候補の探索に基づく枠組においては投票手法をそのまま利用することはできない．なぜならば，同枠組においては上位 n 件の解候補が見つかると検索をそこで終了してしまうので，文書集合全体における解候補についての正確な頻度情報が得られないためである．そこで，投票手法の一つの近似として，以下に述べる疑似投票 (pseudo voting) 手法が導入されている．まず，上位 n 件の解が必要な場合には，**表層表現の異なる**解候補が n 個見つかるまで探索を継続していることに注意されたい．そのため，探索の過程において，すでにスコア計算の終了状態に至っている解候補と全くおなじ表層表現を持つが別の解候補が新たに終了状態に至ることが有り得る．よって，探索の過程において終了状態に至ったすべての解候補を記録することで，解候補の頻度情報を部分的に利用できる．本論文では，解候補 AC に対する疑似投票スコア $S^v(AC, L_q)$ を次のように定義する．

$$S^v(AC, L_q) = (\log_{10}(freq(AC, AnsList)) + 1) \cdot \max_{L_i} S(AC, L_i, L_q) \quad (2)$$

ここで，$AnsList$ は n-best 探索において終了状態に至った解候補のリスト，$freq(x, L)$ を L における x の頻度とする．また，本論文では，この疑似投票スコアを**最終スコア** (final score) と呼ぶことにする．

村田らの研究 [8] によると，上記の投票スコアの計算方法は他の投票スコアの計算方法と同等程度の性能を持つ．

3.3 Web 検索結果中の snippet による Web 文書の利用

文書検索エンジンを Web 検索エンジンに置き換えることにより，基本となる QA システムでも容易に Web 文書を活用できるようになる．しかし，数百もの Web 文書をダウンロードすることは，非常に時間がかかる処理である．この問題に対し，相良ら [11] は Web 検索エンジンによる短い抜粋出力である snippet が Web QA の情報源として有効であることを実験により示している．Katz ら [5] もまた解候補の生成に snippet を使用している．我々も応答時間の短縮のために同様の方法論を採用する．

図 2に基本となる Web QA システムを示す．QA エンジンとしては，前節までに述べた基本となる QA システムが使用されており，情報源として Web 検索エンジンが出力する snippet が用いられる．便宜上，疑似投票を行なう以前の，解候補と原スコアを得る質問応答処理の部分を**主要質問応答処理部** (Core QA part) と呼ぶことにする．

この図において，ラッパープログラム (wrapper program) は個別の検索エンジンの API (Applica-

tion Program Interface) と基本となる QA エンジンとの間に存在するプロトコルの差異を吸収，調整するためのプログラムである．このプログラムはまず最初に主要質問応答処理部からキーワードのリストを受けとり，これを検索要求としてある検索エンジンに入力する．次に，検索エンジンから得られた検索結果から snippet を取りだし，それらを文書集合として主要質問応答処理部に返す．最終的に，主要質問応答処理部ならびに疑似投票部が質問応答処理を行なう．

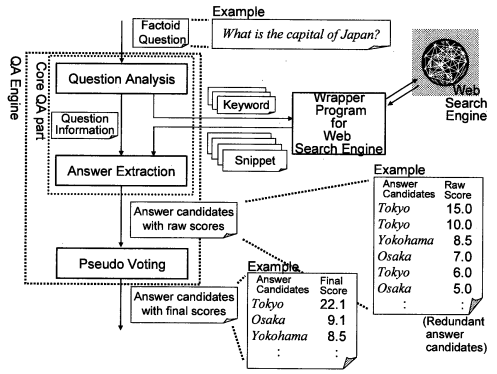

図 2: 基本となる Web QA システム

4 異なる Web 検索エンジンの出力結果を組合せる手法

異なる Web 検索エンジンの出力結果を組合せる手法として以下の 3 手法を検討する．

手法 A: 各検索エンジンから得られた文書集合 (snippet の集合) をそのまま併合し，それを 1 つの QA エンジンに送る．

手法 B: 各検索エンジンから得られた文書集合をそれぞれ個別の主要質問応答処理部 (図 2の Core QA part) に送り，疑似投票を行なう前に，得られた解候補 (原スコアを有する) の各リストを併合する．その後に，式 (2) を使って疑似投票を行なう．次の手法 C とは異なり疑似投票は 1 回だけ行なわれる．

手法 C: 各検索エンジンから得られた文書集合をそれぞれ個別の QA エンジンに送り，疑似投票まで行なった結果として得られた解候補 (最終スコアを有する) の各リストを併合する．併合結果の中においても同じ表層表現を持つ複数の解候補が得られることがあるので，再び式 (2) を使って疑似投票を行なう．

図 3，4，5に手法 A，B，C をそれぞれ示す．各図において，併合部 (merger) は，複数の入力から各々データのリストを受けとり，それらを単純に併合してリストを生成する部分プログラムである．

さて，手法 A はベースライン手法であり，質問応答の処理をする**前に**各検索エンジンの出力結果を併合する．節 2で述べたように，これと同等の方法が先行研究でも採用されている．

一方，手法 B と C は我々の提案手法であり，質問応答処理を行なう前に文書集合を併合するのではなく，質問応答処理を個別に行なった**後に**，解候補を併合するという手法である．ここで，手法 B，C

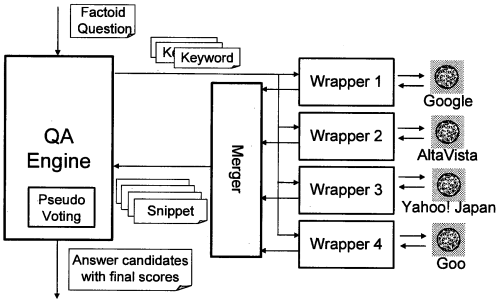

図 3: 複数の Web 検索エンジンを用いる QA システム: 手法 A(ベースライン)

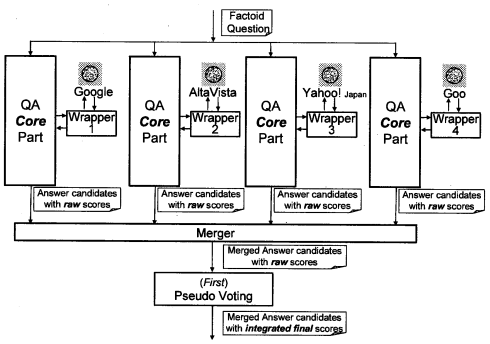

図 4: 複数の Web 検索エンジンを用いる QA システム: 手法 B

が，情報源の多様性の観点において手法 A と異なるということに注意されたい．手法 A と B は，いずれも，疑似投票の前に，まず，Web 検索エンジンが出力する snippet から，原スコアの観点においてより上位の解候補を抽出し，その後に疑似投票を行なうという過程になっているので，一見すると，両者はほとんど同じ解候補を最終的に出力するように見える．しかし，両者の間には次のような差異が存在する．手法 A においては，由来する検索エンジ

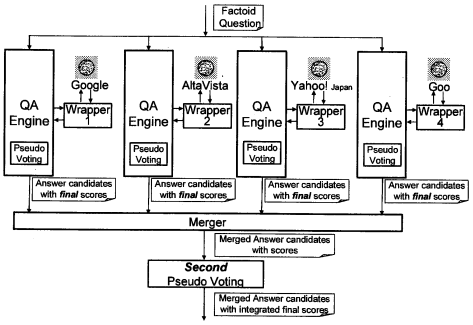

図 5: 複数の Web 検索エンジンを用いる QA システム: 手法 C

ンによる区別をせず，各解候補を同等に扱い，混合された状態でスコアによる順位付けが行なわれるため，最終的に疑似投票にかけられる解候補が，ごく少数の検索エンジンのみに由来し，情報源の利用に偏りが存在する可能性がある．一方で，手法 B ではすべての検索エンジンについて，それに由来する解候補が必ず同数ずつ疑似投票にかけられるので，情報源の多様性が確保されるようになっている．

手法 C も手法 B と同様の状況にあるが，更に，2 段目の疑似投票処理において，より多くの異なる検索エンジンに由来する解候補に対してより高いスコアが与えられる点が手法 B と異なる．

なお，ここでの議論は，解候補に対する確信度が，由来する情報源の多様さによって，ある程度測定できるという我々の仮説に基づいている．この仮説が成り立つのであれば，手法 A よりも，手法 B，C のほうが精度が高いということが期待される．

5 評価実験

前節で述べた 3 つの組合せ手法を評価，検討するために，以下のような質問応答の実験を行なった．特に，利用する Web 検索エンジンの数の効果を調査するために，質問応答における各種設定を同一にした状況において，i) 各 Web 検索エンジンを単独で使用した場合，ii) Web 検索エンジンを 2 つ組合せた場合，iii) 3 つ組合せた場合，iv) 4 つ組合せた場合の各々で実験を行なった．

5.1 評価に用いた質問集合と正解情報

評価に用いる質問文集合としては，NTCIR-3 QAC1 [2] のフォーマルランで用いられた評価用データのうち，質問文集合のみ使用した．同質問集合は人名や地名，製品名といった短い表現が回答となる factoid 型の質問を 200 問有するが，このうち，検索 API のタイムアウト等によりいずれかの QA エンジンが一定時間内に回答を返さなかった 8 問を除外し，計 192 問を評価に用いた．

同評価データには正解となる表現も併せて収録されているが，これは，1998 年ならびに 1999 年の毎日新聞記事を情報源とした時のものであるので，今回の評価には利用できない．そのため，システムが出力した解候補の正解判定については，本評価を行なった 2007 年 9 月～10 月における時事に照らし合わせた上で，第一著者が判断を行なった．

5.2 評価の尺度

求解精度の尺度としては，Mean Reciprocal Rank (MRR) を用いる．Reciprocal Rank(RR) は，システムが出力した順位付き解候補リストにおいて，実際の正解が最初に現れた順位の逆数であり，MRR はその全質問平均である．なお，正解が 5 位以内に現れない時には RR の値は 0 としている．RR ならびに MRR は 0 以上，1 以下の値をとり，値が大きいシステムほど解求精度が高いことになる．

5.3 その他の設定

Web 検索エンジンとしては，以下の商用 Web 検索エンジンを使用した．

1. Google (http://www.Google.co.jp/, Google SOAP Search API),
2. goo (http://www.goo.ne.jp/),
3. AltaVista (http://www.altavista.com/),
4. Yahoo! Japan (http://www.yahoo.co.jp/, Yahoo! JAPAN Developer Network).

すでに述べているとおり，QA エンジンとしては，我々が開発している factoid 型実時間質問応答システムを利用する [7]．本実験で設定した各種パラメタの値を表 1に示す．この設定は，森 [7] において，新聞記事を情報源とした時に最も求解精度が良かったものである．なお，一つのパッセージは隣接する 3 文から構成されている．

ここで注意すべきことは，表 1の設定をいずれの実験においても共通に採用すると，利用する Web 検索エンジンの数を変化させた際に，検索される文書 (snippet) の延べ総数もそれに比例して変化する点である．ここにおいて何をもって公平な比較と考えるかは難しい問題である．もう一つの可能性としては，最終的にシステム全体で参照する文書 (snippet) の延べ総数が同じになるように，表 1のパラメ

表 1: 評価実験における QA エンジンのパラメタ設定

パラメタ名	設定値	説明
a	10	検索すべき解候補の数
d	250	検索すべき文書 (snippet) の数
ppd	5	各文書 (snippet) から抽出するパッセージの最大数
p	30	求解で考慮すべきパッセージの最大数

タ d を調整することが考えられる．しかしながら，森 [7] によると，検索する文書数を増やすことが必ずしも求解精度の向上につながらないことが報告されている．これは，順位の低い文書まで参照することにより，その結果，原スコアは低いものの出現頻度が高い誤った解候補が出現しやすくなるためである．そこで，本論文では単純に，同研究で良いとされた表 1に示すパラメタを用いることにする．

6 実験結果と考察

各 Web 検索エンジンを単独で使用した場合と，Web 検索エンジンを 2 つ，3 つ，4 つと組合せた場合の各々について，前節で述べた質問集合に対する MRR 値を求めた．その結果を図 6に示す．この図

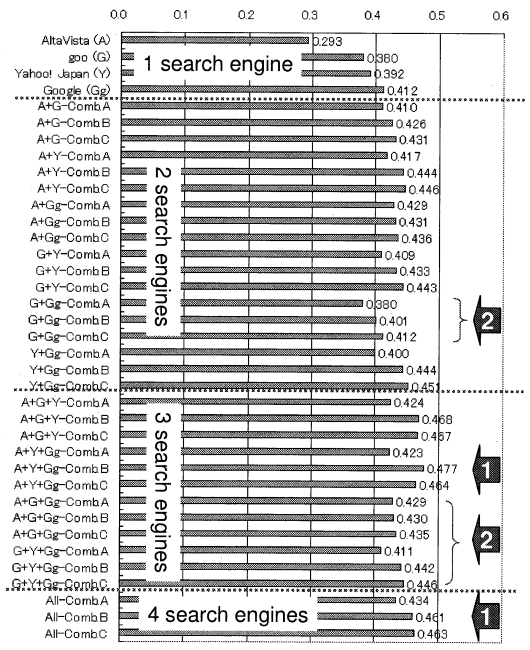

図 6: Web 検索エンジンの各組合せに対する MRR 値

によれば，単独の Web 検索エンジンを使ったシステムよりも Web 検索エンジンを複数組合せるシステムのほうが精度がよくなる傾向にある．特に，質問応答処理を行なう前に文書集合を併合する手法 A に比べ，質問応答処理を個別に行なった後に，解候補を投票処理により併合する手法 B，C のほうが精度がよいことがわかる．一方で，手法 B，C の間の差は顕著ではない．以上より，文書集合を先に併合してから質問応答処理を行なう方法よりも，Web 検索エンジンの出力毎に独立に質問応答処理により解候補を精選した後，それらを投票処理により併合をする方法のほうが精度が良いということがわかる．

また，組合せる Web 検索エンジンの数が増えるに従って，概ね精度が向上することも観察される．一方で，図 6の矢印 1 に示されるように，最も精度が良かった組合せは，すべての Web 検索エンジンを利用した場合ではなく，goo の出力を用いずに，AltaVista，Yahoo! Japan，Google を組合せた場合であることに注意されたい．ここで，図 6の矢印 2 に示される箇所に注目すると，goo と Google の組合せにおいて，他の組合せよりも精度が低い傾向にあることがわかる．この組合せにおいては，Google 単独と同程度以下の精度しか得られていない．その原因を調べるために，同一質問に対して goo ならびに Google が検索結果として何を出力しているかを調べてみた．質問「スイスの公用語は何語ですか．」(QAC1-1140-1) に対する出力を表 2に示す．この

表 2: Google と goo の snippet の比較 (質問が「スイスの公用語は何語ですか．」(QAC1-1140-1) の場合)

rank	Snippet (Google)	Snippet (goo)
1	a 16世紀に、スイス連邦は13の自	☆ スイス連邦の正式名称は4種の公
2	b「スイス」200年近くにわたり中立	a 16世紀に、スイス連邦は13の自
3	c 様々な風景が凝縮された国スイス	b「スイス」200年近くにわたり中立
4	d ドロミテ・ラディン語(ドロミテ語):	c 様々な風景が凝縮された国スイス
5	e「国別情報(国際機関)」■アラビ	d ドロミテ・ラディン語(ドロミテ語):
6	f スイスは4つ公用語を持つ世界で	e「国別情報(国際機関)」■アラビア
7	g 言語．スイスにはドイツ語、フラン	☆ 参照(クリック)〔イタリア語(伊語)
8	h スイスでは、3つの公用語を使用	f スイスは4つ公用語を持つ世界で
9	i だが、スイスとの関連で言えば何	g 言語．スイスにはドイツ語、フラン
10	j 彼が敢えて英語で書かれた拙稿を	h スイスでは、3つの公用語を使用
11	k スイスに住んでいる人はフランス語	i だが、スイスとの関連で言えば何
12	l 言語事情ドイツ語(63．6%)、フ	☆ ドイツ語はスイスの公用語・国語で
13	m スイスは人口七百万人余りの小国	j 彼が敢えて英語で書かれた拙稿を
14	n なかなか難しい条件だったようで	k スイスに住んでいる人はフランス語
15	o Quickbar｜image＝250px｜	l 言語事情ドイツ語(63．6%)、フ
16	p アンドラ公国では公用語、オック語	☆ 年生各学年週4時限履修。2002
17	q スイス － スイス連邦。欧州中部に	m スイスは人口七百万人余りの小国
18	r 時差・時間．－8時間(サマータイ	n なかなか難しい条件だったようで
19	s スイスでは地理的又は歴史的な理	☆ おもしろいことに、ドイツ語と英語(
20	t スイスでは、ドイツ語圏とフランス語	o Quickbar｜image＝250px｜fl

表において，Google の snippet 20 件に対し，異なるアルファベットを ID として付与してある．同様に goo の snippet に対しても Google の snippet と同一の文字列となっている場合に，対応するアルファベットを付与してある．星印は Google の snippet に対応物が存在しない goo の snippet である．表 2によれば Google と goo は同一の snippet を，20 件中

15件と数多く出力していることが分かる．このようなsnippetの同一性は，gooが2003年以降，Googleと提携し，Googleの検索結果を利用しているためだと考えられる．このように数多くの同じsnippetが得られるWeb検索エンジンの組合せにおいては，snippetの表現ならびにそこから抽出される解候補の多様性を担保することができない．例えば，先の質問において，goo，Google，AltaVistaを各々単独で用いた場合に抽出される解候補と，gooとGoogleとの組合せ，ならびに，AltaVistaとGoogleとの組合せにおいて抽出される解候補について調べてみよう．図7にスコアの降順に各上位10件の解候補を示す．

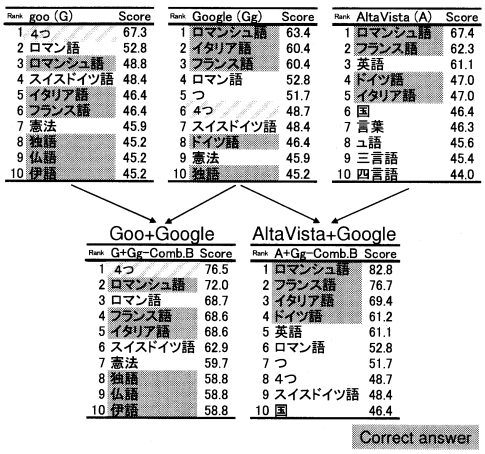

図 7: 各組合せにおける解候補のスコア付けの比較(質問が「スイスの公用語は何語ですか．」(QAC1-1140-1)の場合)

この図に示すとおり，goo単独とGoogle単独の場合の解候補を比較すると，正解，不正解に関わらず同じ文字列が多く出現していることが分かる．一方で，Google単独とAltaVista単独の場合を比較すると，正解である解候補は共通して登場しているものの，不正解の解候補はどちらか一方に出現するのみである．このような状況にあるため，gooとGoogleを組合せた時には，投票処理によって運悪く不正解の解候補のスコアが上昇することもあり，結果としてRR値が相対的に下がることがある．図7左下に示す例では不正解である「4つ」という解候補が1位となってしまっている[1]．一方で，GoogleとAltaVistaの組合せにおいては，図7右下の例のように，正解である解候補のスコアが投票処理によって適切に上昇し，結果として上位4件がいずれも正解となっている．

次に各手法間においてRRの代表値に統計的有意差があるかどうかを調べるために，ウィルコクソンの符号付順位和検定(両側検定)を行なった．RRの代表値に統計的な有意差が見られた手法間の比較の代表例を表3に示す．表中の「p値」はウィルコクソンの符号付順位和検定統計量のp値である．また，表中で「手法X < 手法Y」と記されている時には，手法YによるRRの代表値のほうが手法XによるRRの代表値よりも大きく，なおかつ，その差が統計的に有意であることを示しており，手法Yの方が手法Xより精度が良いことを表す．表3の(1)に示す，「Web検索エンジンの組合せ数に関する比較」の結果を見ると，組合せる検索エンジンの数を増やすほど精度が向上することが統計検定によって確認がなされている．紙面の都合で割愛したが，この表に示した以外の組合せについても，各Web検索エンジンを単独で使用した場合と組合せて使用した場合の間で，統計的有意差が見られるものが数多くあった．次に表3の(2)に示す「組合せ手法に関する比較」の結果を見ると，手法B，Cと手法Aとの間に統計的有意差が見られ，提案した組合せ手法BならびにCの有効性が確認された．また，表3の(3)に示す「gooとGoogleの組合せに関する比較」の結果を見ると，AltaVista，Google，gooの組合せとAltaVista，Google，Yahooの組合せの間に統計的有意差が見られ，検索結果のsnippetが似通っているGoogleとgooを組合せることが精度を劣化させることがわかる．

表 3: 統計検定によりRRの代表値に差が確認された手法間の比較の例とそのp値

RRの代表値に差が確認された手法間の比較	p値
(1)Web検索エンジンの組合せ数に関する比較	
AltaVista (A) < A + Y + Gg (手法B)	2.3×10^{-8}**
Yahoo (Y) < A + Y + Gg (手法B)	0.001**
goo (G) < ALL (手法C)	7.4×10^{-4}**
Google (Gg) < A + Y + Gg (手法B)	0.011*
Y + Gg (手法B) < A + Y + Gg (手法B)	0.018*
A + Gg (手法B) < A + Y + Gg (手法B)	0.018*
G + Gg (手法A) < A + G + Gg (手法A)	0.020*
G + Gg (手法B) < G + Y + Gg (手法B)	0.015*
G + Gg (手法B) < A + G + Gg (手法B)	0.033*
G + Y (手法B) < A + G + Y (手法B)	0.010*
A + G + Gg(手法B) < ALL(手法B)	0.023*
A + G + Gg(手法C) < ALL(手法C)	0.049*
(2)組合せ手法に関する比較	
A + Y + Gg (手法A) < A + Y + Gg (手法B)	0.040*
Y + Gg (手法A) < Y + Gg (手法C)	0.032*
A + G + Y (手法A) < A + G + Y (手法B)	0.030*
A + G + Y (手法A) < A + G + Y (手法C)	0.045*
(3)gooとGoogleの組合せに関する比較	
A + Gg + G (手法B) < A + Gg + Y(手法B)	0.009**

表中 ** は1%水準，* は5%水準で有意差があることを示している．

[1]質問文中にある表現「何語」は「なにご」(何の言語の意味)と解釈すべきであるが，局所的にみると「なんご」(いくつの単語の意味)とも解釈できるので，今回評価に利用したシステムでは，「数」も質問の型としている．

7 おわりに

本論文では，Web 質問応答において，異なる Web 検索エンジンを組合せることによる効果について調査した．具体的には，異なる Web 検索エンジンから得られた検索結果を組合せるタイミングに注目し，異なる 3 手法を比較検討した．評価実験より，質問応答処理の前に検索結果を併合する従来手法よりも，QA エンジンによる個別の質問応答処理の後に，抽出された解候補を組合せるほうが効果があることが示された．また，組合せる Web 検索エンジンの数を増やすほど精度が向上する傾向にあることが観察されたが，同じ snippet を出力する Web 検索エンジンを組合せた時には，snippet の表現の多様性が確保されずに，求解精度が劣化することも確認された．

今後の課題としては，更に精度の良い質問応答を実現するために，a) Web 検索エンジンの出力のより効果的な組合せ手法，b) snippet における表現の多様性を積極的に引き出すために，Web 検索エンジンに入力する検索要求を生成する手法を複数用意して組合せる方法，などについて検討したい．

謝辞

本研究で使用している質問セットは NTCIR-3 QAC1 の成果の一部です．NTCIR-3 QAC1 のタスクオーガナイザならび関係各位に感謝をいたします．

なお，本研究の一部は文部省科学研究費補助金特定領域研究「情報爆発 IT 基盤」(課題番号 No.18049031, 19024033) によるものである．

参考文献

[1] Charles L.A. Clarke, Gordon V. Cormack, and Thomas R. Lynam. Exploiting redundancy in question answering. In *Proceedings of SIGIR '01: the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 358–365, 2001.

[2] Jun'ichi Fukumoto, Tsuneaki Kato, and Fumito Masui. Question Answering Challenge (QAC-1) — Question answering evaluation at NTCIR Workshop 3 —. In *Working Notes of the Third NTCIR Workshop meeting - Part IV: Question Answering Challenge (QAC1)*, pp. 1–6, 2002.

[3] IREX 実行委員会（編）. IREX ワークショップ予稿集. IREX 実行委員会, 1999.

[4] B. Katz, M. Bilotti, S. Felshin, A. Fernandes, W. Hildebrandt, R. Katzir, J. Lin, D. Loreto, G. Marton, F. Mora, and O. Uzuner. Answering multiple questions on a topic from heterogeneous resources. In *Proceedings of TREC 2004*, 2004.

[5] B. Katz, G. Marton, G. Borchardt, A. Brownell, S. Felshin, D. Loreto, J. Louis-Rosenberg, B. Lu, F. Mora, S. Stiller, O. Uzuner, and A. Wilcox. External knowledge sources for question answering. In *Proceedings of TREC 2005*, 2005.

[6] Jimmy Lin and Boris Katz. Question answering techniques for the world wide web, 2003. (Tutorial presentation at The 11th Conference of EACL (EACL-2003)).

[7] Tatsunori Mori. Japanese question-answering system using A* search and its improvement. *ACM Transactions on Asian Language Information Processing (TALIP)*, Vol. 4, No. 3, pp. 280–304, 2005.

[8] Masaki Murata, Masao Utiyama, and Hitoshi Isahara. Use of multiple documents as evidence with decreased adding in a japanese question-answering system. *Journal of Natural Language Processing*, Vol. 12, No. 2, pp. 209–247, March 2005.

[9] Dragomir R. Radev, Weiguo Fan, Hong Qi, Harris Wu, and Amardeep Grewal. Probabilistic question answering on the web. *Journal of the American Society for Information Science and Technology*, Vol. 56, No. 3, March 2005.

[10] Jinxi Xu, Ana Licuanan, and Ralph Weischedel. TREC2003 QA at BBN: Answering definitional questions. In *Proceedings of the twelfth Text Retrieval Conferene (TREC 2003)*, 2003.

[11] 相良春樹, 森辰則, 中川裕志. 質問応答に対する知識源としての web 検索エンジンの snippet の有効性. 言語処理学会第 12 回年次大会発表論文集, pp. 316–319, 3 月 2006.