

# 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

相良 春樹

森 辰則

中川 裕志

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 東京大学情報基盤センター

E-mail : {estrella, mori}@forest.eis.ynu.ac.jp, nakagawa@dl.itc.u-tokyo.ac.jp

## 1 はじめに

近年の計算機性能の向上と、世界規模のネットワークの拡大により、電子化されたさまざまな情報が、即時に得られるようになったが、同時に、ユーザに対する情報過負荷が問題となっている。そこで、ユーザの負担を軽減し、情報の価値を高めるために、自然言語で与えた質問に対して、その回答を大量の文書から発見する質問応答システム(Question Answering System : QA システム)が提案されている。

現在、一般的な質問応答システムでは、知識源として手元にある文書群(ローカルな文書群)を用いているが、文書の範囲や数が限られていることなどから、システムの解答できる質問が当時の時事に限られる、というような制約があった。そこで、代替となる知識源として Web が注目されており、Web 文書を知識源とする質問応答システム(Web QA)が研究され始めている[1]。しかし、Web はローカルな文書群とは異なる特徴を持つため、Web を知識源とする場合、それに対応したシステムを構築する必要がある。

Web 上には膨大な量の文書が存在し、1つの事柄について異なる文書の中で複数回記述される。Web QA には、このような表層的な冗長性を統計的な観点から捕捉し、候補の選定に利用するシステムが多い[2]。正解が数多く出現している知識源を用いた方が、システムが正解を発見しやすいという報告もなされている[3]。Web 文書が豊富であることの大きな要因は著者が世界中に数多く存在するという事実であるが、そのことは記述様式が統一されていないという問題も生じさせる。これが原因で言語的な解析処理が精度良くなされず、質問応答においてシステムが正解を発見しにくいのではないかと懸念がある。

また、通常、Web 検索エンジンを独自に用意することは現実的でないので、既存の Web 検索エンジンを利用する。しかし、知識源となる個々の Web ページの HTML ファイルをダウンロードすると時間がかかるので、これを短縮するためにさまざまな工夫がされる。その中でも、一番簡便な方法は、Web 検索エンジンによる抜粋出力(Snippet)を用いることである。しかし、Snippet は個々のページの一部を、断片的に抜き出したものなので、文章が短く、実際に質問応答の知識源として有効であるのかが未だ明らかではない。

そこで、本研究では、NTCIR-3 QAC-1[4]のテストセットを用いて、実際に Web 検索エンジンより Snippet を取得し、関連する語の統計量の分析を行うことで、質問応答に関する Snippet の性質を調査する。また、Snippet を知識源とする質問応答システムを実装し、新聞記事を知識源としたシステムと比較検討を行う。

## 2 Snippet の分析

### 2.1 Snippet の取得

質問文を解析して抽出された名詞等の内容語を本稿ではキーワードと呼ぶ。このキーワードから構成される検索要求(クエリ)によって、Web 検索エンジン AltaVista[5]から Snippet を取得する。ただし、外部の検索エンジンを調整することはできないので、検索エンジンの性質に適応した工夫をクエリに施すことによって、検索精度の向上を図る必要がある。AltaVista には、利用者の言語環境が特定できない場合、漢字のみで構成されたクエリを用いると中国語で記述された Web ページを多数出力するなどの現象が見られる。日本語の文書のみを検索したい場合には、平仮名などの日本語固有の表現をクエリに加える必要がある。そこで、質問文から得られた文節を、OR 演算子によりクエリに追加した。さらに、質問応答の知識源としては、ある程度の量の Snippet を確保したい。そこで、検索件数が一定の値を超えるまで AND 条件のキーワードを OR 検索の条件に降格し、クエリを自動的に作り直す手法を用いた。

上記の AltaVista に対応したクエリ生成手法を用いて、実際の Snippet を取得し、各種統計量を分析する。調査対象の質問文として、NTCIR-3 QAC-1 のテストセット全 200 問から最初の 50 問を用いた。なお、Snippet の取得は平成 17 年 10 月 6 日に行われ、Web 検索エンジンが出力したすべての Snippet を採用した。

### 2.2 Snippet に関する各種統計量の分析

質問応答の知識源としての Snippet の有効性を、統計量の分析によって検討する。

- 1) 上位 n 件の Snippet 中に正解が現れた問題数(50 問中) 表 1 に結果を示す。

表 1 : Snippet に正解が現れた問題数(50 問中)

Snippet	100 件	300 件	500 件	1000 件
正解が現れた問題数	37 問	41 問	41 問	42 問

上位 1000 件の Snippet 中で正解が現れなかったのは 50 問中 8 問であった。これらの質問を調べると、尋ねていることが、スポーツ用語の略称などの特殊なものであったり、時々刻々と変化する現象のある時点における状態であったりと、元々 Web 文書から見つかりづらい事柄を答とする質問であった。このように、Web 文書の特徴から、Web QA が苦手とする質問が少なからず存在することがわかる。また、長い質問文においては、解を絞り込むのに役立つ、解とは直接関係のないキーワード

が多くなる。これらのキーワードを含む検索要求により検索を行ったために、期待するような文書が得られない場合があった。

2) 上位 n 件の Snippet 中に現れる一問当たりの正解の数

表 2 に結果(平均値)を示す。なお、括弧 [ ] 内の数字は、正解が 1 回以上現れた 42 問に限定したときの平均値を表す。

表 2 : Snippet 中に現れる一問当たりの正解の数[回/問]

Snippet	100 件	300 件	500 件	1000 件
延べ数	12.60 [17.03]	35.00 [42.68]	49.28 [60.10]	75.02 [89.31]
異なり数	2.06 [2.78]	3.16 [3.85]	3.56 [4.34]	4.36 [5.19]

この表より、Snippet1000 件中出现する正解の延べ数が 75.02、異なり数が 4.36 と、比較的多くの正解が出現することがわかった。しかし、問によっては正解の出現回数が少ないこともあるので、解を選択する指標として、解候補の冗長性を期待し頻度だけを用いる手法は誤った候補を解とってしまう可能性がある。

3) 質問文から抽出されたキーワードが同一 Snippet 中、同一文中に出現する回数と正解の有無との関係

表 3 に、同一 Snippet 中、同一文中に出現するキーワードの数(平均値)を示す。調査は、正解が出現する場合と、正解が出現しない場合に分けて行った。なお、斜字体の数字は、質問文に含まれるキーワードの異なり数を母数としたときの、現れたキーワードの異なり数の割合を表す。上段は質問毎に平均を計算した後、その値を質問に亘って平均した、いわゆるマクロ平均である。一方、下段は質問の区別とは無関係に全ての平均を計算した、いわゆるマイクロ平均である。

表 3 : 同一 Snippet 中、同一文中に出現するキーワードの数[回]

全 Snippet, 文の場合(50 問)					
同一 Snippet 中			同一文中		
延べ数	異なり数		延べ数	異なり数	
5.34	3.53	<i>0.68</i>	1.17	1.04	<i>0.21</i>
		<i>0.64</i>			<i>0.19</i>
正解が出現する Snippet, 文の場合(42 問)					
同一 Snippet 中			同一文中		
延べ数	異なり数		延べ数	異なり数	
7.98	4.59	<i>0.87</i>	3.36	2.69	<i>0.54</i>
		<i>0.83</i>			<i>0.49</i>
正解が出現しない Snippet, 文の場合(50 問)					
同一 Snippet 中			同一文中		
延べ数	異なり数		延べ数	異なり数	
5.16	3.45	<i>0.67</i>	1.13	1.01	<i>0.2</i>
		<i>0.63</i>			<i>0.18</i>

この表より、正解が出現する Snippet や文においては、正解が出現しないものと比べて、キーワードの延べ数、異なり数ともに多く出現することがわかる。逆に言えば、キーワードが多く出現する Snippet や文には、正解が出現しやすい。既存の質

問応答システムにおいて、パッセージ検索等の順位付けにキーワードの延べ数や異なり数を用いるものがあるが、Web QA においても、上記の性質を使って、キーワードの延べ数や異なり数に基づき Snippet や文の順位付けを行うことができる。

4) Snippet 中の正解とキーワードとの間の距離

表 4 に結果(平均値)を示す。なお、距離は正解と最も近いキーワードとの間で計っておりに現れる数を示し、隣接関係が距離 1 で表される。なお、文節数において、正解とキーワードが同一文節内で出現した場合は、距離を 0 とした。

表 4 : 正解と最も近いキーワードとの距離

文字数	形態素数	文節数
4.75	2.91	1.01

この表より、文字数を単位として約 5、形態素数を単位として約 3 となり、正解はキーワードから比較的近い位置に現れることがわかる。また、文節数を単位としてみると距離が約 1 となり、隣接する 2 つの文節内に正解とキーワードがそれぞれ出現する機会が多いことがわかる。

5) 正解が、質問の焦点や質問文中で疑問詞に最も近いキーワードと同一 Snippet 中、同一文中で現れる数

一般に、質問の焦点や質問文中で疑問詞に最も近いキーワードが正解と密接な関係にある。質問の焦点(Question Focus : QF)とは、質問の解の範囲を限定するキーワードであり、疑問詞の表す質問の型の下位概念である。例えば、「2006 年の冬季オリンピックが開催される国はどこですか」という質問であれば、解を限定する「国」が QF である。

表 5 に、QF や質問文中で疑問詞と最も近いキーワードが、正解とともに現れる問に限定したときの結果(平均値)を示す。なお、斜字体の数字は、各質問において得られた Snippet 群に現れる全正解数を母数としたときの、QF や質問文中で疑問詞に最も近いキーワードが同一 Snippet 中、同一文中で共起した正解の数の割合であり、上段はマクロ平均値、下段はマイクロ平均値である。

表 5 : QF や質問文中で疑問詞に最も近いキーワードと同一 Snippet 中、同一文中で現れる正解の数[回]

QF (QF, 正解がともに現れる 26 問)		質問文中で疑問詞に最も近い キーワード(正解が現れる 42 問)			
同一 Snippet 中	同一文中	同一 Snippet 中	同一文中		
112.92	<i>0.90</i>	64.38	<i>0.67</i>	83.98	<i>0.90</i>
	<i>0.95</i>		<i>0.68</i>		<i>0.94</i>
				53.88	<i>0.68</i>
					<i>0.74</i>

この表によれば約 90%の正解が、同一 Snippet 中で QF や質問文中で疑問詞に最も近いキーワードなど特別なキーワードと一緒に出現している。このことから、QF や質問文中で疑問詞に最も近いキーワードを含まない Snippet を、解候補の探索範囲から除外することが、対象を絞るという意味で有効な手段であると予測される。

6) 同一 Snippet 中, 同一文中で正解と一緒に出現する, 質問の焦点や質問文中で疑問詞に最も近いキーワードの数

表 6 に, QF や質問文中で疑問詞に最も近いキーワードが, 正解とともに現れる間に限定したときの結果(平均値)を示す. なお, 斜字体の数字は, 各質問において得られた Snippet 群に現れる全ての QF や質問文中で疑問詞に最も近いキーワードの数を母数としたときの, 正解が同一 Snippet 中, 同一文中で共起した割合であり, 上段はマクロ平均値, 下段はマイクロ平均値である.

表 6: 同一 Snippet 中, 同一文中で正解と一緒に出現する QF や質問文中で疑問詞に最も近いキーワードの数[回]

QF (QF, 正解がともに現れる 26 問)				質問文中で疑問詞に最も近い キーワード(正解が現れる 42 問)			
同一 Snippet 中		同一文中		同一 Snippet 中		同一文中	
130.19	<i>0.18</i>	68.73	<i>0.09</i>	114.83	<i>0.14</i>	65.24	<i>0.08</i>
	<i>0.19</i>		<i>0.10</i>		<i>0.14</i>		<i>0.08</i>

この表によると, QF や質問文中で疑問詞と最も近いキーワードの全出現に対する, 正解と同一 Snippet 中で出現する割合は 20%未満で, これらの特別なキーワードを手がかりにして正解を探すという戦略は有効に働かないと考えられる.

### 3 指標の提案

#### 3.1 Snippet の再順位付け

質問に関連する語の統計量の分析から, 正解の出現する Snippet には, ほかのものよりキーワードが多く含まれるということが明らかとなった. そこで, 質問文に含まれるキーワードの延べ数, 異なり数, QF の延べ数と Snippet 中での出現の有無, 質問文中で疑問詞に最も近いキーワードの延べ数と Snippet 中での出現の有無などの情報を組合せることにより Web 検索エンジンから得られた Snippet の再順位付けを行い, 上位に正解を集められるかを調査した.

幾つかの組み合わせを調査した結果, キーワードの延べ数と異なり数の積を順位付けの指標とし, その値の降順に Snippet を並べ替える手法が一番効果的であることがわかった. これは, 延べ数, 異なり数の両方においてキーワードが頻出する Snippet が正解を含みやすいということである. 表 7 に, 同手法で再順位付けする前と後の一問当たりの正解の出現回数(平均値), 正解が現れる問題数を示す. また, 他の組み合わせの検討結果によれば, キーワードの異なり数よりも延べ数を用いたときの方が, 結果が良くなった. おそらく, 質問文に含まれるキーワードの異なり数が元々多くないので, 全ての Snippet において, 出現するキーワードの異なり数があまり変わらないことが原因であると考えられる.

表 7: 再順位付けの前と後での正解の出現

	正解が現れる問題数			正解の出現回数[回/問]		
	Snippet	100 件	300 件	500 件	100 件	300 件
前	37 問	41 問	41 問	17.03	42.68	60.1
後	41 問	41 問	42 問	30.83	60.17	73.26

#### 3.2 再順位付けに伴う変化を用いた語の重要度

前節で, 質問文に含まれるキーワードの出現回数などにより, Snippet の再順位付けを行った結果, 正解の語句を含む Snippet をランキングの上位に集めることができることを述べた. この結果から, 再順位付けにより上位に集められた Snippet に含まれる語は, 質問と密接に関連し, 質問応答において重要な役割を担う語であると考えられる. さらに, その中には正解も数多く含まれていると考えられる. そこで, 再順位付け前後の Snippet の順位の変化を用いて, この重要語(形態素)を判定する手法を提案する.

##### 3.2.1 提案手法

まず, Snippet に含まれている文を形態素に分解する. そして, 形態素ごとに以下の値を求め, これらの指標の再順位付け前後での変化を語の重要度とし, この値が正になったものを質問に関連する重要語と判定する.

i) 指定した上限順位以内での形態素の出現回数

質問文に含まれるキーワードによって Snippet が再順位付けされるので, 質問に関連する語は, 上位に集中すると考えられる. したがって, 特に関連の強い語は, 上限順位以内での出現回数が, 大きく増加すると考えられる.

ii) 注目する形態素を含む Snippet の平均順位

再順位付けで順位が上がった Snippet に出現する語は, 質問に関連があると考えられる. そこで, 各形態素について, それを含む Snippet の順位をその形態素の順位と定義し, その平均値を指標とする. ただし, 必ずしも平均順位は上がらないので, 偏差値を用いて, 概ね値が[0,1]の区間に含まれるように正規化する.

iii) 注目する形態素を含む Snippet の順位分散

注目する形態素が出現する Snippet が, 再順位付けで上位の方に押し集められ, 順位分散が小さくなった場合, その語は質問に関連があると考えられる. ただし, 必ずしも分散が小さくならないので, ii)の場合と同様に偏差値を用いて正規化する.

##### 3.2.2 いくつかの利用方法の提案

この手法によって得られた重要語(形態素)は, 以下のような場面で利用できると思われる.

a) キーワードと重要語の両方を使って, 再度, 再順位付けを行う. 例えば, キーワードと重要語の両方を含む Snippet を重要視し, キーワードだけが出現する Snippet よりも上位となるようなスコア付けを行う.

b) 重要語を解候補の語句の一部として, 解抽出の手がかりとする. この指標を使って Snippet に含まれる形態素を順位付けしたときに, 正解の語句に含まれる形態素の最高順位(最も上位の形態素の順位を問に亘って平均した順位)が最も上位だったのは, 次の条件のときであった. 再順位付けには, キーワードの延べ数で再順位付けした後, 同順位のもの異なる

り数で順位付けする手法を用いる。語の重要度の判定には、再順位付け前後の、全 Snippet 数の上位 3 分の 1 以内での語の出現回数之差と、順位分散の比との積を指標にし降順に並べる手法を用いる。以上の条件のとき、一問当たりの上位 3 分の 1 の Snippet 中に出現する全ての語(形態素)の数、重要度の値が正になった重要語の数、その中で正解の一部となっている語の最高順位(平均値)を表 8 に示す。なお、括弧 [ ] 内の数字は、正解が 1 回以上現れた 42 問に限定したときの平均値を表す。この表より、Snippet 中に正解が現れる 42 問において、上位 3 分の 1 の Snippet 中に出現する約 3000 語の中で質問に関連する重要語と判定されたのは 651 語で、そのうち正解の語句の一部は 53 位で得られた。

表 8 : 重要語の判定(一問当たり)

全ての語の数(異なり数)	重要語の数(異なり数)	正解の一部の最高順位
3035.60 [2952.17]	672.16 [651.29]	169.84 [53.10]

### 3.3 問題点

キーワードの延べ数と異なり数の両方が多い Snippet に正解が出現する可能性が高いことがわかっているため、そういった Snippet が再順位付けで上位に集まるのが好ましい。しかし、質問文に含まれるキーワードの異なり数が元々少ないため、延べ数の影響が強く、複数あるキーワードのうち、延べ数が多い特定のキーワードによって Snippet の順位が引き上げられる場合が見られた。その結果、前節で提案した手法によって判定した重要語の上位に、特定のキーワードだけに関連するようものが多く見られた。

## 4 Web QA システムの実装と評価

これまで Web 検索エンジンの Snippet を質問応答の知識源として用いることの有効性を確認してきたが、本節では Web QA の実装例を紹介し、精度の評価を行う。

我々は、既存の Web 検索システムに対する検索を代行し、Snippet を取得する検索代行サーバを実装し、既存の質問応答システム[6]と組み合わせることにより Web QA システムを構築した[7]。なお、前節で提案した Snippet の再順位付けは未実装である。ここでは質問文として NTCIR-3 QAC-1 のテストセット全 200 問を用い、知識源となる Snippet は上位 250 件を使用し、MRR(Mean Reciprocal Rank. 最上位正解(5 位以内)の順位の数値を全質問に亘って平均したもの)によってシステムを評価する。Web 検索エンジンとしては AltaVista に加えて goo[8]も使用した。また、比較のために、NTCIR-3 QAC-1 で使用した 98 年、99 年の毎日新聞の記事を知識源としたときのシステムの精度も評価した。このシステムは質問応答エンジンには同じものを用いており、文書検索部が異なるだけであるため、上記 Web QA システムに対する精度の上限の目安となる。結果を表 9 に示す。なお、括弧 ( ) 内の数字は、新聞記事を知識源としたときの MRR に対する比率を表す。

この表より、新聞記事を知識源としたときに比べて、Snippet を知識源としても比較的よく回答できていることがわかる。我々のシステムでは解候補のスコアに、文字 2-gram の頻度、キ

ーワードの出現や頻度、文節の依存構造の類似度、固有表現抽出に基づく質問型的一致など、様々な水準の詳細度や頑健性を持つ言語処理に基づく複数の指標を導入しており、この点が上記の結果に寄与していると考えられる。

この結果からも、質問応答の知識源として、Snippet は十分有効であることがわかる。

表 9 : 各知識源を用いたときの MRR

	新聞記事	AltaVista	goo
MRR	0.512 (100%)	0.412 (80.5%)	0.459 (89.6%)

## 5 まとめ

本研究では、Web 検索エンジンが出力する Snippet の質問応答の知識源としての有効性を探るべく、NTCIR-3 QAC-1 のテストセットを用いて、実際に Web 検索エンジンから Snippet を取得し、さまざまな観点から分析を行った。それにより、今まで明らかでなかった、質問応答の知識源としてみた Snippet の性質が明らかとなった。また、Snippet 中のキーワードと正解の語の共起の関係から、Snippet の再順位付けを行い、さらに、再順位付けに伴う語の統計量の変化から、質問に関連のある重要語の判定の手法について提案した。

また、Snippet を取得する検索代行プログラムを実装し、既存の質問応答システムと組み合わせることにより Web QA システムを実装し、評価を行った。その結果から、質問応答の知識源として、Snippet が十分有効であることが示された。

今後の課題としては、再順位付けに使う指標を精緻化するとともに、今回明らかとなった質問応答に関する Snippet の性質を考慮して、我々の質問応答システムに Web 文書向けの調整を行うことが挙げられる。

## 参考文献

- [1] Jimmy Lin and Boris Katz. Question Answering Techniques for the World Wide Web. Tutorial presentation at EACL-2003 (2003)
- [2] Dell Zhang, Wee Sun Lee. Web Based Pattern Mining and Matching Approach to Question Answering. TREC 2002 (2002)
- [3] Marc Light, Gideon S. Mann, Ellen Riloff, Eric Breck. Analyses for Elucidating Current Question Answering Technology. Natural Language Engineering. (2001)
- [4] Jun'ichi Fukumoto, Tsuneaki Kato, Fumito Masui. Question Answering Challenge (QAC-1): An Evaluation of Question Answering Task at NTCIR Workshop 3. Proceedings of the Third NTCIR Workshop (2003)
- [5] AltaVista. <http://www.altavista.com/>
- [6] Tatsunori Mori. Japanese Question-answering System Using A\* Search and Its Improvement. ACM Transactions on Asian Language Information Processing (to appear)
- [7] 質問応答システム(WWW). <http://www.forest.eis.ynu.ac.jp/Forest/ja/services.html>
- [8] goo. <http://www.goo.ne.jp/>