社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－NL－178 (13)
2007／3／29

# 質問応答システムにおける知識源選択規則の自動獲得の有効性について

亀山 恵祐† 荒木 健治† 木村 泰知‡

†北海道大学大学院 情報科学研究科 ‡小樽商科大学 商学部 社会情報学科

本研究では，質問応答システムの応答検索部において，学習データから得られたルールを適用し検索対象となる最適な知識源を選択する手法を提案する．従来手法では，質問応答における知識源として新聞記事やWWWなど単一の知識源を検索することにより応答を生成していた．提案手法では，知識源として新聞記事とWWWの双方を用い，さらに新聞記事に関しては年代別に細分化し，学習ルールに基づき選択することにより精度の向上を試みる．実験の結果，NTCIRのQACタスクに用いられた質問文において，新聞記事やWWWの単一の知識源を用いた場合に比べ，細分化した知識源を使い分けることによって正解数が増大し，本手法の有効性が確認された．

# Effectiveness of Automatic Acquisition of Knowledge-Source Selection-Rules for a Question Answering System

Keisuke KAMEYAMA† Kenji ARAKI† Yasutomo KIMURA‡

†Graduate School of Information Science and Technology Hokkaido University
‡Department of Information and Management Science Otaru University of Commerce

We propose a method for increasing correct responses by proper choice of the knowledge source using selection-rules at the Answer-Retrieval section within our Question-Answering System. We use both the newspaper and WWW as the knowledge sources. And we divided the newspaper corpus into separate parts according to the publishing year. Through the experiments with the questions for NTCIR-QAC tasks, we confirmed the increase of correct responses when one knowledge source is chosen compared with the case where they are used separately.

## 1. はじめに

現在，自然言語処理の中でも情報検索の分野は日々進化しており，多くの検索エンジンが開発され，人々のニーズも高まってきている．その一つの要因として，日本においては情報インフラの整備が進み，日本人の実に7割弱の人がインターネットを使用できる環境下にあることが挙げられる [1]．また情報技術が向上したことにより，安価で高性能の情報機器が手軽に利用できる様になったことも一つの要因と考えられる．だが，こうした情報化社会が進む一方で，それゆえに起こる問題もある．それは，大量の情報を誰でも簡単に利用できるために起こる，ユーザにとっての情報過多という問題である．一つの目的で情報を検索していても，その条件に合致するものが大量にあるために，ユーザは情報の取捨選択を迫られる．ある程度使い慣れているユーザは情報の取捨選択を日常的に行っているためにあまり負担とはならないだろうが，初心者にとっては，この情報の取捨選択が容易ではなく，負担となっている．

そこで，自然言語処理の分野では，上記のような情報の取捨選択にかかるユーザの負担を減らすべく，質問応答システムが研究されている．この質問応答システムとは，ユーザが自然言語で入力した質問文に対して，大量の文書群である対象知識源からその応答となる文や語を検索し，提示するものである．今日，質問応答システムに対しては様々な試みがなされている．その中でも1990年代に米国で開始された

TREC(Text REtrieval Conference) [2] は広く一般的に知られており，ここから数多くの情報検索技術が発展している．また国内でもNTCIR(NII/NACSIS Test Collection for Information Retrieval) [3] という同様の評価型ワークショップが開催されている．こちらは検索対象が日本語であり，日本語特有の言語処理に関する問題を取り扱うものとして，これまでに6度開催されている．

質問応答システムに関する研究は数多く行われており，意味的制約を用いて質問タイプを同定し，質問文や抽出した文を構文解析して解を推定する研究 [4]，質問文解析・応答検索・抽出・応答出力をすべて学習データからの自動学習による研究 [5]，パターンマッチングや構文解析を用いる研究 [6]，回答候補前後の文字列と，質問の類似度の尺度として構文構造の類似を利用する研究 [7] などがある．

本研究の焦点は，システムが検索の対象とする知識源である．一般的に質問応答の知識源としては，新聞記事やWWWが代表的である．しかし，それらは非常に大量の文書であり，そこから目的の情報を正確に得ることは容易ではないと考えられる．そこで，それら知識源の細分化を行い，使い分けることを考える．知識源を細分化することで，検索範囲が狭められるために検索時間を短縮することができ，さらに詳しく調べることができる．いわば，質問応答システムが応答が含まれると考えられる文書を抽出した後，さらに絞り込んで出力する語を決定するように，まず応答が含まれると考えられる知識源を選択して，そこからさらに通常の質問応答システムの処理を行うことで，より多くの正解を得ようとするものである．

簡単に新聞記事とWWWの特徴を述べる．新聞記事は表現が正確であり，それゆえにコンピュータによる解析が行いやすい．しかし新聞記事は全ての事柄を網羅し掲載できるわけではなく，記事にならない事柄もある．一方，WWWにおける文書では，個人がその文書の著者になれるため，どんな記事も存在しうる．その一方で，表現が多様であり，口語表現など，コンピュータにとっては解析しづらい文書があることも多い．このように，新聞記事とWWWの文書にはそれぞれ長所と短所が存在する．そこで，これら2つの知識源の長所を活かし，短所を補うようなシステムを構築したものが本研究で提案する質問応答システムである．これまでは，知識源選択に関連した研究はほとんど行われていない．本研究は，こうしたそれぞれの知識源の長所や短所を考慮した上で，質問に対して最適な知識源を自動的に決定し，応答を生成することを目的としている．

本研究では，実験を通して提案手法の評価を行い，本システムの有効性について考察を行う．以下，第2章でシステム処理過程について述べ，第3章では知識源についての解説を行う．第4章では，知識源選択規則について説明し，第5章では，提案手法の有効性を確認した実験結果と考察について述べる．最後に第6章で本研究のまとめを示す．

## 2. システム概要

本研究で提案するシステムの概要を図1に示す．本システムは我々がNTCIR-5 QAC-3用に構築したシステム [8] を改良したもので，以下の6つの処理部から構成される．

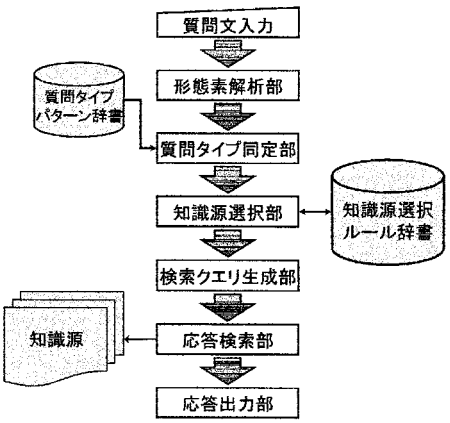


図 1: システム概要図

**形態素解析部** ユーザがキーボードから自然言語で入力した質問文に対して形態素解析ツール『茶筌』 [9] を用いて単語分割を行い，品詞情報を付与

**質問タイプ同定部** 形態素解析部の結果を受け質問タイプパターン辞書とのマッチングを図り質問タイプを同定

**知識源選択部** 質問文の形態素解析結果と質問タイプを基に，システムが適切であると判断した知識源を選択

**検索クエリ生成部** 形態素解析部の結果と選択した知識源を基に，質問文中の語を組み合わせてクエリを生成

**応答検索部** 知識源選択部で選択した知識源から，検索クエリ生成部で生成したクエリを含む文書を抽出

**応答出力部** 固有表現抽出ツール『NExT』 [10] を用いて，抽出した文書の中から応答として相応しいと考えられる語を出力

## 3. 知識源

本システムが検索対象とする知識源として NTCIR の QAC タスクで用意された新聞記事，毎日新聞と読売新聞の 1998 年から 2001 年までの各 4 年分，計 8 年分を利用した．新聞記事のデータ検索の即時性と利便性を高めるために，全文検索システム Namazu [11] によるインデキシングを行った．インデキシングを行うことにより，対象となる文書に対し，キーワードを入力することで，そのキーワードを含む新聞記事を即座に出力することができるようになる．ここで，両新聞をひとつにまとめた，それぞれの年に対する Namazu によるインデキシングの結果を表 1 に示す．

表 1: 新聞記事 8 年分のインデキシング結果

| 年 | 記事数 | 総キーワード数 | ファイル容量 (KB) |
|---|---|---|---|
| 1998 | 248,517 | 1,620,629 | 345,191 |
| 1999 | 347,541 | 2,067,685 | 469,831 |
| 2000 | 405,915 | 2,282,627 | 532,943 |
| 2001 | 452,484 | 2,442,979 | 545,240 |
| 合計 | 1,454,457 | 6,901,070 | 1,893,205 |

もう一つ，WWW からの知識源として Google [13] による検索結果の Snippet を用いた．通常 WWW を知識源として用いるためには，自ら検索エンジンを開発するか，既存の検索エンジンを用いることになる．しかし，自分で検索エンジンを開発するには，非常に大きな労力と技術が必要であるので，本システムの補助ツールとしては自ら検索エンジンの開発を行わず，既存の検索エンジンを用いるものとした．また，既存の検索エンジンで検索された個々のサイトを一つずつ解析していくのは，対象となる文書が大量になる上，HTML などの解析を行う必要が出てきてしまい，新たな処理が必要となる．その結果，処理時間の増加につながってしまう．そこで本システムでは，検索エンジンによって検索したサイトを解析する時間を短縮させる方法として，相良ら [14] により最も簡便であり，有効であると確認された Web 検索エンジンの抜粋出力である Snippet を利用することとした．この Snippet を用いることにより，WWW から 100 件のサイトを解析してもインデキシングした新聞記事数年分を検索するのと，ほぼ同等の処理時間で応答を返すことが可能となる．

### 3.1 新聞記事

知識源として新聞記事を用いることの長所は，そこに記載されている記事内容はほぼ正確な情報であり信頼できるものである，ということである．これは，社会的に認められた企業が全国的に公表しているものであり，その信頼性は高いものであるといえる．また口語表現はほとんど無く，用いられている文体や表現も統一性があるため，表記の揺れが少なく，計算機にとって解析しやすい文であるともいえる．一方で短所として挙げられるのは，新聞記事が対象とする範囲に制約があり，記事の数も限られているために，あらゆる分野を網羅できるわけではないことがある．さらに出版物であるため，情報が最新のものではない可能性があるといった短所がある．

知識源としての新聞記事の選択肢の幅を広げるために，毎日新聞と読売新聞の区別なく各年ごとで分けた．それを 1998 年，1999 年，2000 年，2001 年の 1 年分，1998-1999 年，1999-2000 年，2000-2001 年の 2 年分，1998-2000 年，1999-2001 年の 3 年分，1998-2001 年の 4 年分，全部で 10 通りの新聞記事による知識源を用意した．これらをそれぞれ全文検索システム Namazu [11] でインデキシングを行い，即座に検索結果を得られるようにした．新聞記事の特徴は前述した様に，WWW とは対照的に内容が限定されていることが，質問応答システムにとっては有効に働く場合が多くある．

### 3.2 WWW

新聞記事に比べて WWW では，情報が常に更新されており，対象となり得る文書も新聞記事に比べて非常に大量に存在するという長所がある．また最近では，ウィキペディア [12] 等，管理体制の整った信頼性が高い文書も数多く存在し，そうした文書は信頼性の高いコーパスとして用いることが可能である．その一方で短所としては，誰にでも WWW 上のデータの著者となることができるので，Web 文書の信頼性が新聞記事に比べると低く，誤った文書も十分に存在し得ることを配慮しなければならない．また，個々人が書く文書には口語的な表現が含まれることもあり，システムで解析するのは容易ではない可能性があるといったことが挙げられる．

今回は，既存の Web 検索エンジン Google で検索した際の抜粋出力 Snippet を利用する．検索エンジンを Google としたのは，世界で最も広く使用されている検索エンジンであり，プログラミングをする上でも組み込みやすいためである．

## 4. 知識源選択規則

はじめに挙げたように，本研究はそれぞれの知識源の長所や短所を考慮した上で，質問に対して最適な知識源を自動獲得することを目的としている．そこで，学習により知識源を選択できるようにシステムの構築を行った．

まず，質問文データとして過去のNTCIRのQACタスク [15] の質問文を用いる．本研究で使用するものは，QAC1，QAC2タスク1・2が各200問，およびQAC3のReference1の質問文が360問の合計960問である．

### 4.1 学習ルール

知識源選択規則を獲得するための学習データとして，質問文と質問タイプ，更にその正解のセットを利用する．質問文に対して正解を得られた場合，その質問文を形態素解析し，その結果を他の正解が得られた質問文と比較を行い，共通部ルールと差異部ルールを得る．

知識源選択部における知識源選択ルール辞書に存在するルールは，学習による共通部ルールと差異部ルールである．共通部ルールとは，質問文が正解を得た知識源ごとに質問文を分け，同じ知識源から正解が得られた質問文の共通部である．例えば，いくつかの質問文において1999年の新聞記事から正解が得られた場合，それらの質問文と質問タイプを比較し，共通部分を獲得する．差異ルールとは，WWWから正解が得られた質問文と新聞記事から正解が得られた質問文を比較して，どちらか一方にのみ正解が得られた場合，その差異部分を差異ルールとして抽出する．その際，どの知識源から得られたルールなのかも同時に保持する．これらのルールにより，新たな質問文に対して，それぞれのルールにマッチする場合に，そのルールが保持する知識源に重みを加えていく．そして最も値の高かった知識源を最適な知識源と判断し適用する．

共通部ルールの獲得の例を図2に示す．

Q1とQ2の2つの質問文が共に2001年から正解が得られたとする．その際，これら2つの質問文の茶筌による解析結果と質問タイプを比較し，それらの共通部分を2001年の知識源に対する共通部ルールとして獲得する．

## 5. 実験方法

本稿で提案した知識源選択規則の自動獲得を利用したシステムの有効性を確認するための実験を行った．選択規則を獲得するために学習データと評

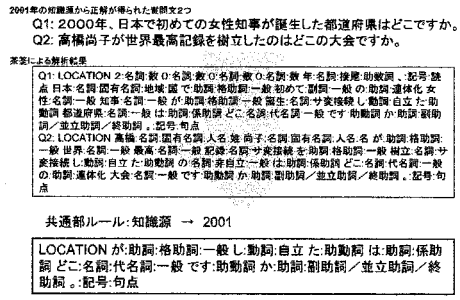


図 2: 共通部ルールの獲得例

価データに分類するが，本実験では，QACのタスクごとに質問データを分類した．ここで扱うQACのタスクは4つあるので，学習データとして3つのタスクを使用し，残り一つのタスクを評価データとして使用した．すなわち，例えばQAC1のタスクを評価データとして用いた場合，学習データはQAC1Task1および2そしてQAC3の質問文が使われるということである．各QACタスクの質問文において正解を得られた質問文 (図2における異なり質問数) が，学習および評価データの対象となる．

### 5.1 実験結果

まず，本システムが対象とする11の知識源，それぞれを使用した際に得られる正解数を表2に示す．ここで，異なり質問数とは，本システムにおいて，対象とする11の知識源のいずれかで正解が得られた質問文数である．表中の太字で下線が引かれている箇所は，新聞記事中で最も正解が多く得られた知識源である．また，単独の知識源で最も正解数が多く得られるのはWWW(表中の太字二重下線)であることが分かる．

次に提案手法による実験結果を表3に示す．表3

表 3: 提案手法による質問に対する正解数

| タスク | QAC1 | QAC2 Task1 | QAC2 Task2 | QAC3 |
|---|---|---|---|---|
| 異なり質問数 | 38 | 44 | 61 | 42 |
| 提案手法 | 26 (68.4) | 29 (65.9) | 43 (70.5) | 22 (52.4) |
| WWWのみ | 22 (57.9) | 21 (47.7) | 34 (60.7) | 21 (47.6) |

括弧内は異なり質問数に対する割合 [%]

表 2: 各知識源ごとの正解数

| 知識源 | QAC1 | | QAC2Task1 | | QAC2Task2 | | QAC3 | |
|---|---|---|---|---|---|---|---|---|
| 全質問数 | 200 | | 200 | | 200 | | 360 | |
| 1998 | 16 | (8.0) | **16** | **(8.0)** | 18 | (9.0) | 4 | (1.1) |
| 1999 | 9 | (4.5) | 13 | (6.5) | 16 | (8.0) | 8 | (4.0) |
| 2000 | 10 | (5.5) | 10 | (5.5) | 8 | (4.0) | 10 | (2.8) |
| 2001 | 14 | (7.0) | 5 | (2.5) | 11 | (5.5) | 11 | (3.1) |
| 1998-1999 | 14 | (7.0) | **16** | **(8.0)** | **24** | **(12.0)** | 3 | (0.8) |
| 1999-2000 | 12 | (6.0) | 14 | (7.0) | 16 | (8.0) | 13 | (3.6) |
| 2000-2001 | 14 | (7.0) | 8 | (4.0) | 9 | (4.5) | **15** | **(4.2)** |
| 1998-2000 | 15 | (7.5) | **16** | **(8.0)** | 23 | (11.5) | 7 | (1.9) |
| 1999-2001 | 13 | (6.5) | 11 | (5.5) | 16 | (8.0) | 12 | (3.3) |
| 1998-2001 | **17** | **(8.5)** | 13 | (6.5) | 21 | (10.5) | 12 | (3.3) |
| WWW | **22** | **(11.0)** | **21** | **(10.5)** | **34** | **(17.0)** | **20** | **(5.6)** |
| 異なり質問数 | 38 | (11.7) | 44 | (22.0) | 61 | (30.5) | 42 | (11.7) |

括弧内は全質問数に対する割合 [%]

より，共通部ルールと差異部ルールを用いることにより，単独の知識源を用いるよりも，複数の知識源を用いる方がよりよい結果が得られたことが示された．

## 6. 考察

知識源選択規則の適用による成功例と失敗例をそれぞれ図 3，図 4 に示す．

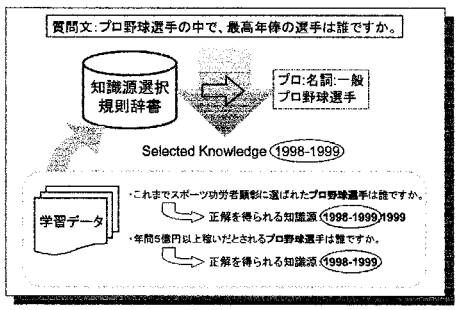


図 3: 選択規則適用の成功例

成功した場合には，図 3 にあるように関連のありそうな同じような質問文が学習データ中に存在することが多い．例に挙げたように,『プロ野球選手の中で、最高年棒の選手は誰ですか。』という質問に対して，学習データ内に『これまでスポーツ功労者顕彰に選ばれたプロ野球選手は誰ですか。』,『年間 5 億円以上稼いだとされるプロ野球選手は誰ですか。』といったように，同記事内に書かれていそうな内容についての質問文がある．これは，知識源選択規則辞書内の差異部ルールにおいて，“プロ:名詞:一般” “プロ野球選手”が含まれる質問文に対して，1998 年と 1999 年の 2 年分の知識源から検索することが最も正解を得やすい，とシステムが判断したためである．

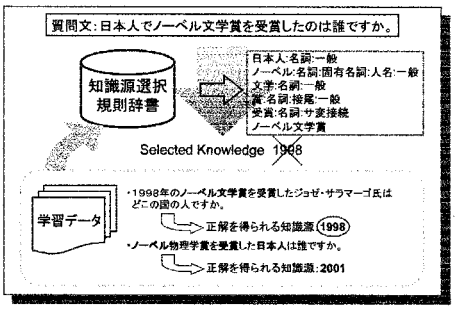


図 4: 選択規則適用の失敗例

一方，失敗した場合には，図 4 にあるように関連はあるが，別の記事に書かれていそうな場合がある．例の場合,『日本人でノーベル文学賞を受賞したのは誰ですか。』という質問文に対して，“日本人:

名詞:一般”, “ノーベル:名詞:固有名詞:人名:一般”, “文学:名詞:一般”, “賞:名詞:接尾:一般”, “受賞:名詞:サ変接続”, “ノーベル文学賞”といったルールとマッチする学習データ中の質問文が正解を得た知識源のうち，最も尤度の高かった1998年から検索したが，正解しなかった．この場合，“ノーベル文学賞”と“日本人”がキーワードとなっており，どちらも満たす質問文が学習データに存在しなかったためにルールが得られず，どちらかが質問文中にある質問文のうち尤度計算の結果で1998年の新聞記事から検索をしたが，正解が得られなかった．

以上のように，差異部ルールを用いることにより，学習データ内の類似している質問文の結果を反映させた知識源の選択をするようになる．しかし，類似した質問が複数ある場合には，間違った知識源から検索してしまうこともある．これは，更なる知識源選択のための学習データの解析が必要である．しかし，更なる学習データの解析によって，コンピュータの性能が同じであると仮定すれば，質問文を入力してから応答が返ってくるまでの時間が長くなってしまう恐れがあり，応答時間についても考慮する変更が必要であると考えられる．

## 7. まとめ

本研究では，質問応答システムにおいて知識源の細分化による新たな正解数増加の可能性を考慮し，それらの知識源に対し最適なものを自動的に選択できる質問応答システムの提案を行った．過去のNTCIRのQACタスクに参加した際のシステムに改良を加え，さらに知識源の選択を自動的にできるよう実験システムを構築した．知識源選択規則を獲得するために質問文を解析し，その質問文に対し正解が得られた知識源とのペアを学習データとして用い，それらの共通部と差異部をルールとして獲得した．実験では，細分化した知識源を選択することでより正解が得られることの有効性を確認する実験を行った．学習によって獲得した共通部ルールと差異部ルールを評価データに適用し，知識源選択規則の有効性を確認した．

今後の本研究の発展としては，新聞記事だけではなくWWWにおける知識源の選択肢を増加させることが考えられる．現在はGoogleによる検索結果のSnippetのみを利用しているが，検索先を特定の情報提供サイトに限定したものや，ブログやニュースサイトだけに限定したものなどに特化することでWWWからの知識源の選択肢を増大させることができる．また，これらの多様な知識源を学習によって自動判別するためには，検索時間を考慮して行う必要がある．これらを今後の課題としたい．

## 参考文献


[1] Internet World Stats. http://www.internetworldstats.com/

[2] TREC. http://trec.nist.gov/

[3] NTCIR. http://research.nii.ac.jp/ntcir/

[4] 村田 真樹, 内山 将夫, 井佐原 均. “類似度に基づく推論を用いた質問応答システム”情報処理学会 自然言語処理研究会, NL-135, pp.181–188, (2000).

[5] 佐々木 裕, 磯崎 秀樹, 鈴木 潤, 国領 弘治, 平尾 努, 賀沢 秀人, 前田 英作. “SVMを用いた学習型質問応答システムSAIQA-II”情報処理学会論文誌, Vol.45, No.2, pp.635–646, (2004).

[6] Eduard Hovy, Ulf Hermjakob, Chin-Yew Lin. “The Use of External Knowledge in Factoid QA”, Proceedings of TREC-11, pp.644–652, (2001).

[7] Tetsuro Takahashi, Kozo Nawata, Shinya Kouda, Kentaro Inui. “Seeking Answers by Structural Matching and Paraphrasing”, Proceedings of NTCIR-3 Workshop Meeting, pp.87–94, (2003).

[8] Yasutomo Kimura, Kenji Ishida, Hirotaka Imaoka, Fumito Masui, Marcin Skowron, Rafal Rzepka, Kenji Araki. “Three Systems and One Verifier - HOKUM's Participation in QAC3 of NTCIR-5”, Proceedings of NTCIR-5 Workshop Meeting, pp.402–408, (2005).

[9] 松本 祐治, 北内 啓, 山下 達雄, 平野 善隆, 松田 寛, 浅原 正幸. “形態素解析システム茶筌”, http://chasen.aist-nara.ac.jp.

[10] 固有表現抽出ツール:NExT. http://www.ai.info.mie-u.ac.jp/ next/next.html

[11] 全文検索システム:Namazu. http://www.namazu.org/

[12] フリー百科事典:ウィキペディア. http://ja.wikipedia.org/

[13] Google. http://www.google.co.jp/

[14] 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, (2006).

[15] QAC. http://www.nlp.is.ritsumei.ac.jp/qac/