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

2004－ＦＩ－ 76　(18)
2004－ＮＬ－163　(18)
2004／9／17

質問応答における常識的な解の選択と期待効用に基づく回答群の決定

秋葉友良[†]　藤井敦[‡]　伊藤克亘[*]

[†] 豊橋技術科学大学　[‡] 筑波大学　[*] 名古屋大学

akiba@cl.ics.tut.ac.jp

概要　オープンドメイン質問応答に関する新しい手法を提案する。第 1 に、コーパス中に遍在する「常識」を利用した回答候補選択手法を提案する。従来の固有表現抽出を用いたカテゴリ一致検査による回答候補選択手法に対して、事前知識構築を必要とせず比較的単純な仕組みで実現可能という実装上の利点に加え、カテゴリ粒度を質問に応じて動的に決定するため高精度の検査が可能という性能向上の点でも利点がある。第 2 に、リスト型質問に対して期待効用最大化原理に基づく回答群選択手法を提案する。従来ヒューリスティクス的手法がほとんどであった回答群選択の問題に対し、理論的な意味付けを与え、性能向上も達成した。2 つの手法に対し、評価実験により有効性を示す。

Question Answering using Common Sense Knowledge latent in Corpora and Utility Maximization Principle

Tomoyosi AKIBA[†]　Atsushi FUJII[‡]　Katunobu ITOU[*]

[†] Toyohashi University of Technology, [‡] University of Tsukuba, [*]Nagoya University

Abstract　In this paper, we propose two new methods for open-domain question answering. First, we use knowledge resembling “common sense” for question answering purposes. For example, the length of a runway in an airport must be a few kilometers, but a few centimeters. In practice, we use specific types of information latent in document collections to verify the correctness of each answer candidate. Second, we use the utility maximization principle to determine the appropriate number of answers for a list question. We estimate the expected value of the evaluation score, on the basis of the probability scores for multiple answer candidates. We show the effectiveness of our methods by means of experiments.

1　まえがき

米国 NIST の TREC-8(1999)[15] や日本での NTCIR-3(2002) から大規模な評価が行われている質問応答は、自然言語の質問文による検索質問について、組織化されていないオープンドメイン文書集合から直接の答となる部分のみを抽出する、精度重視の情報検索技術である。本稿では、筆者らが NTCIR-4 の Question Answering Challenge(QAC) 2[4] に参加した質問応答システム [1] で採用した、質問応答の 2 つの新しい手法について述べる。

質問応答では、関連文書から回答候補を抽出する際に、回答候補が現れるコンテキストと質問文との類似性の基準と共に、回答候補自体の意味的な制約を用いることが多い。これらの手法では、人手で記述した規則 [10]、既存のシソーラス [12]、事前獲得した知識 [3] などを用いて、回答候補の意味的妥当性を検査する。一方、特にオープンドメインの文書集合を対象とした質問応答では、意味的な制約を規則で厳密／網羅的に記述することが難しい場合もある。

質問「愛知県の常滑沖にできる新しい空港の滑走路は開港時どのくらいですか。」では滑走路の何の属性について訪ねているのか、計算機が質問文を文字通り解析しただけでは判断できない。しかし、人間が見れば滑走路の話題として、長さや本数が適切であることは明らかである。また「滑走路」の長さは、数キロメートル程度が普通で、数メートルなどは解答となり得ないことも明らかである。人間が持つこのような知識は「常識」と呼ばれ、日常生活における種々の推論に意識すること無く普通に用いられている。

「常識」は人間が経験した多くの事例を背景とする。したがって、事例を多く含む新聞記事などの大規模なコーパス中には「常識」が眠っていると考えられ、それを知識源として利用すれば、質問応答の性能向上に役立つはずである。本稿では、回答候補の妥当性を判定するためにコーパスを直接知識源として用いる新規手法を提案する。提案手法はコーパスを一般知識源として用いるので、質問応答の対象以外のコーパスも利用可能で、かつコーパスをなるべく多く使うほど効果が期待できる。

第 2 の手法は、リスト型質問 (list question) と呼ばれる問題定義に関する手法である。リスト型質問では、正解数が与えられていない質問 (正解がない場合も含む) に対し、システムは正解だけを網羅的かつ重複無く答えることが要求される。1 つの質問に対しシステムは回答をいくつ出力してもよいが、誤った回答を出力すると減点となる。提案手法は、ある回答群を出力したときの効用関数の期待値を比較することで、適切な回答群を選択する手法である。

本稿の構成は次の通りである。2 節では、本稿における質問応答の問題設定を行う。続く 3 節では第 1 の手法、4 節では第 2 の手法を、それぞれ説明し、評価実験の結果を示す。5 節で結論を述べる。

2　質問応答の問題設定と従来手法

質問応答は、「質問文解析」「関連文書の検索」「回答候補抽出」「回答候補の順位付け」といった一連のプロセスとしてとらえることが多い。しかし、本稿では次のように探索の問題として考える。

質問応答　質問文 q と文書集合 D が与えられていると

き、D 中の全ての部分文字列の出現の集合 $S = \{(d, p_s, p_f) | d \in D, p_s p_f$は d 中の位置$, p_s < p_f\}$ について、q の回答としての $a \in S$ の相応しさを表す評価関数 $L(a|q)$ によって、回答 $\hat{a} = \mathrm{argmax}_{a \in S} L(a|q)$ を求める。

この問題設定は、正解を一つ見つける問題で、TREC の factoid question や NTCIR QAC の Subtask1 に対応する。この問題設定に従えば、質問応答の研究課題は、(A) 適切な評価関数 $L(a|q)$ を設計すること、(B) 効率よく $L(a|q)$ を最大化する回答を求めること、と考えることができる。本稿では、(A) に焦点を当てる。

従来の質問応答システムにおいて、評価関数 L は次の 2 つの基準を組み合わせて構成することが多い。

a. 回答候補のコンテキストに関する基準

b. 回答候補自体に関する基準

a は、回答候補前後の (回答候補を含まない) 文字列 (文、パッセージ、段落) と、質問の類似度の基準である。類似の尺度としては、共通の単語を用いる手法を基本として、構文構造の類似を利用するものなどが提案されている [13][16]。また、コンテキストとしてどの程度の範囲 (パッセージ) を用いるかの検討も行われている [14]。

b の回答候補自体の基準には、回答候補の意味に関する制約が用いられる。質問文解析によって正解の意味カテゴリを推定、検索対象文書集合を固有表現抽出することによって得られる回答候補の意味カテゴリとの一致を調べる手法が、多くのシステムで採用されている。一般に、固有表現抽出においてより細かいカテゴリ分類を用いると、回答候補のより高精度なチェックが可能になるため、質問応答の性能が向上する。例えば、文献 [10] では独自に設定した 62 のカテゴリを、文献 [8] では階層的に分類した 189 のカテゴリを用いており、それぞれ QAC1,QAC2 の Subtask1 でトップの成績を達成している。一方、固有表現抽出を用いる手法には次のような問題点がある。

- 質問文解析や固有表現抽出に必要な知識の構築コストが高い。カテゴリを詳細化するほど、構築コストは高くなる。機械学習による知識の自動獲得も試みられている [9] が、大量の教師付データを用意する必要があるためやはり高価である。
- 質問応答の性能が、質問文解析や固有表現抽出の精度に依存する。一般により細かいカテゴリを利用するほど抽出の精度は低くなるため、カテゴリの過度の詳細化は、逆に質問応答の性能を引き下げてしまうこともある。

本稿で提案する第 1 の手法は、b の回答候補自体の基準について、固有表現抽出や事前に定義した意味カテゴリを用いる代りに、事前知識構築しない素のコーパスだけを用いて候補の意味制約を検査する手法である。(3 節)

一方、NTCIR QAC の Subtask2,3 や、2003 年度からの TREC QA Track の一部では、リスト型質問がタスクの一つとして採用されている。リスト型質問の問題設定は、上記の問題設定において、評価関数 L を回答群 A の相応しさを表す $\mathbf{L}(A|q)$ に置き換え、これを最大化する回答群 $\hat{A} = \mathrm{argmax}_{A \in 2^S} \mathbf{L}(A|q)$ を求める問題として定義することができる。しかし、従来の手法では、回答群の評価関数 $\mathbf{L}$ の代りに、各回答候補の評価関数 L と、ヒューリスティクスによって回答数を決定する手法がほとんどであった。本稿で提案する第 2 の手法は評価関数 $\mathbf{L}$ として、リスト型質問の評価に用いられる効用関数 (F 値) の期待値を用い、回答群を選択する手法である。回答群選択の問題に対し、理論的な意味付けを与えた点に特徴がある。(4 節)

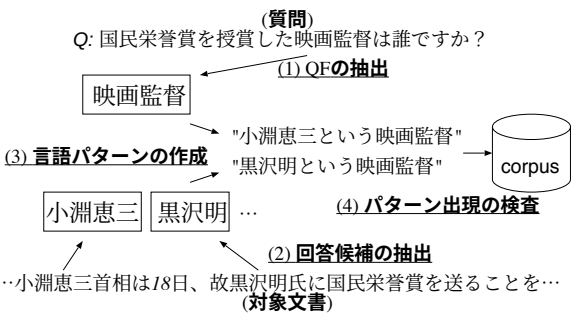

図 1: 提案手法の処理過程

3 常識的な解の選択

質問応答の質問文には、回答に期待する意味的な制約を直接表す表現がそのまま現れることが多い。例えば、質問「２０００年のＮＨＫ大河ドラマは何ですか。」には、正解が「ＮＨＫ大河ドラマ」のインスタンスであることが示されている。このような、正解に関する制約を直接表す質問の中心的な表現を、本稿では Question Focus (以下、QF) と呼ぶ。

質問文から QF を抽出し、回答候補との意味的な関係の有無を検査することによって、回答候補を選択する手がかりとすることができる[1]。QF が「ＮＨＫ大河ドラマ」である場合、「新選組」は「ＮＨＫ大河ドラマ」のインスタンスなので回答候補として適切だが、「おしん」(ＮＨＫ連続テレビ小説のインスタンス) は適切ではない。

従来、このようなクラス-インスタンス関係、あるいは上位下位関係、を調べるためには、既存のシソーラス (WordNet、EDR 辞書、等) あるいは事前に自動獲得した知識 [3] が用いられてきた。提案手法は、このような事前に体系化した知識を利用せず、組織化されていないコーパス (プレーン・テキスト) を知識として利用する。コーパスは一般知識源として利用するため、質問応答の対象文書に限らない。関係判定の再現率を上げるために、なるべく大規模なコーパスを用いるのが望ましい。2 つの語 (QF と回答候補) の意味的関係の検査は、コーパス中にその 2 つの語を含む特定の言語パターンが現れるかどうかを調べることで実現する。例えば、「…新選組などのＮＨＫ大河ドラマは…」なる表現がコーパス中に現れていることで、「新選組」は「ＮＨＫ大河ドラマ」のインスタンスであることがわかる。

提案手法による意味関係検査の処理を図 1 に示す。以下、3.1 節では質問文から QF を抽出する処理、3.2 節では対象文書から回答候補を抽出する処理、3.3 節では QF と回答候補の間の意味関係を検査する処理について詳しく説明する。

3.1 QF の抽出

まず、質問文から QF を抽出する。QF と認められる文字列は、単独の単語から、修飾句まで含めた大きな句まで、様々な単位が考えられる。例えば、質問「久石譲が音楽を担当した宮崎駿監督の映画は何ですか。」について、「映画」「宮崎駿監督の映画」「久石譲が音楽を担当した宮崎駿監督の映画」などの単位が考えられる。一般に、大きな単位を QF とすると、QF の特定性が高くなるためより高精度な検査が可能になるが、文書集合に現れる頻度が少なくなり再現率は下がる。また、句のような構文的な多義性を伴う単位を自動抽出すると、手法が解析の精度に左右されることになる。以上を考慮にいれて、本稿では次の単位を QF と認定した。

[1]質問によっては QF が無いこともあり、その場合には提案手法は適用できない。例えば、質問「小淵首相は就任直後、何と酷評されましたか。」には QF が存在しない。

- 名詞または隣接する名詞の連続。
- 単位の大きさに選択の余地がある場合は、最も大きい単位を選択。
- 「もの」「名前」「人物」「場所」など、疑問詞から分かる回答のカテゴリと同等あるいはより低い特定性を持つ候補は選択しない。

QF の抽出には、質問文の形態素列 (表層文字列 (SS)、基本型 (BS)、品詞 (POS) などを素性として持つ構造体の列) に対する、人手で記述した正規表現パターンを用いた。例えば、次のようなパターンにより括弧の部分を QF として抽出する[2]。

```
({ POS=>名詞 }+) "の" "名前" "は" "何"
```

3.2 回答候補の抽出

2 節の問題設定に従えば、対象文書集合中のすべての部分文字列を回答候補として抽出すればよい。しかし、部分文字列全体から成る空間 S は巨大なので、一般に質問応答システムでは、関連文書検索、文解析 (形態素解析等)、固有表現抽出などによって見込みのある候補だけを抽出し、探索空間の絞り込みを行う。

本稿でも同様の近似手法を用いる。まず、質問文から抽出した索引語集合を検索質問として関連文書検索を行う。検索された文書を形態素解析し、全ての自立語と、正規表現パターンで指定した単語列 (数値表現、日付、姓名、カタカナ文字列、など)、括弧内の文字列を候補として抽出する。さらに、各候補は質問文とのコンテキストの類似性、質問中の疑問詞から判別できるカテゴリとの一致度、などでスコアを与え、上位 N 個の候補を回答候補とする。

3.3 意味関係の検査

QF と回答候補 (AC) の意味的関係の有無を次の方法で検査する。まず、QF と AC を両方含む文集合を、“QF and AC” を検索質問とした文書検索および検索された文書からの文の切り出し処理によって求める。次に、人手で作成した QF と AC を変数として含む単語列の正規表現パターンを用いて、特定のパターンの発見を試みる。

TREC における factoid question や、NTCIR QAC1,2 では、回答は名称か数量表現のどちらかとなる。どちらになるかは、質問文解析によってほぼ推定可能である。例えば、質問中の疑問詞が「誰」「どこ」「どの」ならば回答は名称に、「いつ」「いくら」「どれくらい」では数量表現となる。一方、「何色ですか」と問われた場合は、名称 (色の種類) か数量表現 (色の数) か決定できない。このような場合は、両方の可能性を同時に処理し、最終的に求まる評価関数の値によって回答を選ぶ。

以下の節では、回答が名称と数量表現になる場合それぞれについて、提案手法の処理と利用した正規表現パターンについて説明する。

3.3.1 名称の検査

QF と回答候補の間の上位下位関係の有無を、それらの関係を表す形態素列パターンがコーパスに現れるかどうかで検査する。利用したパターンの例を図 2 に示す。図中の QF と AC は変数であり、実行時はそれぞれ質問文から得られる QF と対象文書に現れる回答候補に置き換えて使用する。各パターンには、信頼性によってスコアが付与されており、検出した最も高いスコアを採用する。

```
AC {SS=>"という",POS=>連語} QF
AC {SS=>"など",POS=>副助詞} {SS=>"の",POS=>助詞}? QF
AC {SS=>"の",POS=>助 詞} {SS=>"よ う",POS=>助 動 詞 語 幹 } {SS=>"な",POS=>助動詞} QF
AC {SS=>"以外",POS=>非自立名詞} {SS=>"の",POS=>助詞} QF
QF {SS=>"「",POS=>括弧開} AC {SS=>"」",POS=>括弧閉}
QF {SS=>"・",POS=>記号} AC {POS=>付属語}
AC {SS=>"・",POS=>記号} QF {POS=>付属語}
QF AC {POS=>付属語}
AC QF {POS=>付属語}
QF {SS=>"の",POS=>助詞} AC {POS=>付属語}
```

図 2: 上位下位判定の判定パターン

同様のパターンを用いて単語間関係を獲得する手法には、文献 [6] をはじめとして多数提案されている。ただし、先行研究が知識の「獲得」に用いたのに対し、提案手法では関係の有無の「検査」に用いる点が異なる。

QF と正解の間に他の意味関係が認められる場合も考えられるが、本稿では上位下位関係だけを対象とする。例えば、質問「日本が負担している在日米軍駐留経費は別名何と呼ばれていますか。」では、QF を「在日米軍駐留経費」とすると正解との間には同値 (同義語) 関係が認められる[3]。

3.3.2 数量表現の検査

回答候補の数量表現は「数値+単位」のパターンで現れる。「数値」と「単位」それぞれを、以下の方法で妥当性を検査する。ただし、本稿では正解が日付となる場合は対象としない[4]。

単位の検査

回答候補の単位の妥当性を、QF と単位の意味関係を見つけることで検査する。この意味関係には、質問「ＮＨＫ連続テレビ小説の平均視聴率は最高どのくらいですか。」の QF「平均視聴率」と単位「%」のような属性と単位の関係と、質問「愛知県の常滑沖にできる新しい空港の滑走路は開港時どのくらいですか。」の QF「滑走路」と単位「メートル」のような明示されない属性 (例の場合「長さ」) を間にはさんだ対象と単位の関係が含まれる。後者の場合、常識的な属性を暗に推定していることに相当する。

これらの意味関係を、次の正規表現で表したパターンがコーパスに現れるかどうかで検査する。

```
QF {POS=>付属語}* {POS=>数字}+ UNIT
```

ここで、QF と UNIT は変数であり、実際の QF と回答候補中の単位で置換する。例えば、QF「記憶容量」と回答候補「100ＭＢ」からパターン「"記憶容量" {POS=>付属語}* {POS=>数字} "ＭＢ"」が得られ、コーパスにこのパターンを満たす「記憶容量は 650 ＭＢ」という表現が現れることで単位「ＭＢ」が適切であることがわかる。

文献 [11] では、提案手法と同様のパターンを用いて、事前に妥当な QF と単位の組を抽出し、回答候補の選択に用いている。先行研究が知識の「獲得」を行うのに対し、提案手法では関係の有無の「検査」に用いる点が異なる。

数値範囲の検査

数値範囲に関する常識は、質問応答の回答選択に役立つ。例えば、「国立大学の入学金は２０００年度からいくらになると決まりましたか」という質問に対し、「…国立大学の入学金の値上げは１０００円減の２０００円アップで決着した。２０００年度入学者から実施され２７万７０００円に

[2]本稿で用いる正規表現パターンを、次のように表す。括弧 {} は一つの形態素に対応する。その構成要素 ATTR=>VAL は属性 (ATTR) と値 (VAL) の組を表し、形態素が満たすべき制約を示す。形態素のパターンが {SS=>"..."} の場合は単に"..."と表記する。

[3]同義語を問う質問文には特定の表層表現パターンが認められる。今後、そのような質問の判別と、同義語関係を検査するパターンによって、提案手法の考え方で処理可能と思われる。

[4]正解が日付となる場合は、質問文の疑問詞が「いつ」「何日」などとなることで容易に特定可能である。

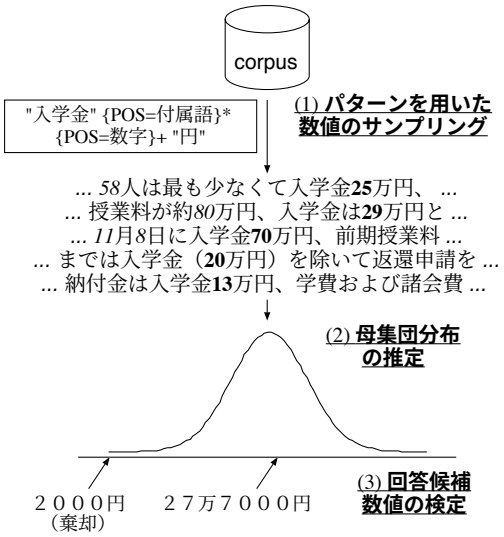

図 3: 数値範囲の検査

なる。…」という記事を見つけたとする。ここから「１０００円」「２０００円」「２７万７０００円」の回答候補が得られるが、３つのうちから正解を選ぶのは難しい。しかし、人間が見れば「入学金」に関する常識的な値から「２７万７０００円」を選ぶことができる。

コーパスに現れる数量表現は、ある話題 (QF) についての常識的な値の事例と考えられる。よって、コーパス中の事例との近さを調べれば、回答候補の数値の尤もらしさが判定できる。「単位の検査」に用いたパターンを用いて、QF と「単位」が共起するコンテキストでの数値の集合を抽出する。ここで抽出した数値集合を、QF に関する数値母集団からランダムサンプリングして得られた標本とみなし、この標本から母集団分布の推定を行う。本稿では、母集団分布として正規分布を仮定した[5]。

求めた分布に対して、回答候補の数値がこの分布からの標本であるという仮説について検定し、回答候補の検査を行う。具体的には、仮説が棄却されないような両側確率 (危険率) を求めて評価関数 L に反映した。図 3 にこの過程を図示する。図の例では、QF「入学金」について数値をサンプリングし、平均約 25 万円の母集団分布を推定している。この分布に対して、回答候補「２７万７０００円」は分布の平均に近いので尤もらしいが、「２０００円」は分布から外れるため適当ではないことがわかる。

文献 [7] では、数値に関する意味カテゴリ毎に常識的な数値範囲を知識として保持し、回答選択に用いている。先行研究が事前に定義したカテゴリを対象に事前獲得した知識を用いるのに対し、提案手法は質問に応じて動的に得た意味カテゴリを対象にコーパスから直接範囲の推定を行う点が異なる。

3.3.3 接尾表現による検査

コーパスから意味関係を検査する以上の手法に加え、回答候補の接尾表現だけを用いた手法を併用する。

回答が名称である場合、回答候補の接尾表現が QF そのものである場合、QF と回答候補が上位下位関係であると見なすことができる。例えば、「天保山」(回答候補) が「山」(QF) のインスタンスであることは、「天保山」の末尾に「山」があることから判定できる。

[5]特定の QF を考えると、正規分布以外の分布が適当と考えられるものも考えられる。例えば、QF「交通事故」の件数はポアソン分布に従うと考えられる。しかし、本稿では QF と確率分布の関係に関する知識が無い場合の近似として、最も一般的な正規分布を仮定した。

表 1: 上位下位関係検査の再現性 (問題数)

QF の有無	関係検査		2 年	13 年
QF あり	成功		69	90
		pattern	62	84
		suffix	22	22
	失敗		49	28
QF なし			31	31
計			149	149

表 2: 単位表現検査の再現性 (問題数)

QF の有無	関係検査		2 年	13 年
QF あり	成功		24	26
		pattern	16	21
		suffix	13	13
	失敗		9	7
QF なし			0	0
計			33	33

回答が数量表現である場合、質問文に回答の単位が明示されている場合があり、これを手がかりとすることができる。例えば、質問「チュニジアの人口は何人ですか。」の「何人」から回答の単位が「人」であることがわかる。

3.4 評価実験

3.4.1 意味関係検査の再現率

提案手法による意味関係検査の再現率を調査した。まず、QAC1 formalrun Task1 の 200 問から QF を人手で抽出した。次に、正解を持つ質問 196 問について、抽出した QF とテストコレクションから得た正解の両方を含む、意味関係の検査に用いる言語パターンのいずれかが、新聞記事に出現しているかを調べた。一つの質問に二つ以上の正解がある場合は、それぞれの正解について調べ、そのうちの一つでも関係が認められた場合に成功とした。意味関係を検査するコーパスは、13 年分の新聞記事 (毎日新聞 11 年分と読売新聞 2 年分) を用いた。

名称

名称を問う質問は、196 問中 149 問であった。3.1 節の定義に従い、正解の上位語にあたる名詞または名詞連続を QF として抽出した。149 問中、QF が抽出できたものは 118 問 (79.2%) であった。抽出した 118 の QF とそれぞれの回答文字列について、上位下位関係があるかどうかを提案手法によって検査した。結果を表 1 に示す。列 “2 年”、“13 年” は、判定に用いた新聞記事のサイズ (年数) を示す。118 問中、パターン検索を用いた判定 (表中の “pattern”) 単独で 84 問 (71.2%)、接尾表現の検査と併用した場合 (“suffix”) で 90 問 (76.3%) の関係が判定できた。

単位

数量を問う質問は 196 問中 47 問、その中で答が日付のものは 14 問であった。日付以外の 33 問すべてについて QF が抽出できた。

抽出した 33 問の QF と、各質問の正解数量表現を構成する単位について、関係の有無を提案手法によって検査した。結果を表 2 に示す。33 問中、パターン検索を用いた判定 (表中の “pattern”) 単独で 21 問 (63.6%)、接尾表現と併用した場合 (“suffix”) で 26 問 (78.8%) の関係が判定できた。

数値

単位が検査できた 21 問中、10 個以上のサンプルが抽出できた 14 問について、数値検査の性能を調査した。正解の数量表現の数値が、サンプルから推定した分布からの標本で

表 3: 正解数値と危険率の関係

危険率	0.5	0.3	0.1	0.05	0.03
棄却される	8	5	2	1	0
問題数 (%)	(57.1)	(35.7)	(14.2)	(7.1)	(0.0)

表 4: QAC1 を対象とした質問応答システムの評価

回答種別	system	MRR	AFM
名称	BASE	0.453	0.316
153 問 (149 問)	+pattern	0.533	0.401
数量表現	BASE	0.475	0.343
47 問	+pattern	0.450	0.332
	+sampling	0.461	0.330
全体	BASE	0.458	0.322
200 問 (196 問)	+pattern	0.513	0.384
	+sampling	0.516	0.384

あるという仮説について両側検定を行った。結果を表 3 に示す。サンプル数は少ないものの、正解との関連がほぼ確認できた。

コーパスのサイズ

提案手法は、パターン検索の対象とするコーパスのサイズを増やすことで、意味関係検査の再現率を改善することができる。新聞記事 2 年分と 13 年分の比較により、その効果を確認することができた。また、本実験で検査に失敗した組の多くは、Web の検索エンジンによってパターンを含む文書を見つけることができ、サイズ増加で再現率をさらに上げることが可能であることを確認した。特に数値の検査手法では、本実験の 13 年分で十分なサンプル得た組は少なく、サイズの増加が望まれる。

3.4.2 質問応答システムの性能評価

提案手法を質問応答システム [1] に実装して、システム全体の性能を調べた。提案手法の意味関係検査は、回答候補それぞれについて文書検索と正規表現パターンマッチングを繰り返すため、実行時の処理コストが大きい。そのため、評価関数 L のうち提案手法以外で求まる値により回答候補を順位付けし、上位 20 候補に対してのみ提案手法を適用するといった、近似手法を用いた。

QAC1 テストコレクション

まず、NTCIR QAC1 テストコレクションを用いて評価を行った。質問文解析における QF の解析精度を Task1,2 の 200 問[6]で調べたところ、QF のある 154 問中 139 問 (90.3%) を正しく抽出し、誤検出は 200 問中 11 問 (5.5%) であった。

システム全体の性能評価の結果を表 4 に示す。列 “MRR”、“AFM” は、それぞれ QAC1 の Subtask1(Mean Reciprocal Rank)、Subask2(Average F-Measure) の結果を表す[7]。行の “BASE” は、意味検査に 3.3.3 節で述べた接尾表現だけを用い、コーパスからのパターン検索を用いない場合の結果、“+pattern” はパターン検索を用いた場合、“+sampling” はさらに数値の統計情報を用いた場合を示す。

名称を問う 153 問 (うち正解のあるもの 149 問) について、Subtask1 の MRR 評価で提案手法により改善が見られた質問は 24 問、逆に悪化した質問は 8 問であった。悪化した質問 8 問を調べたところ、誤った QF の抽出 (1 問)、関係判定の失敗 (2 問)、別候補の沸き出し (5 問)、が原因であった。

表 5: QAC2 を対象とした質問応答システムの評価

system	BASE	+提案手法	+正解 QF
MRR	0.476	0.498	0.522

数量表現を問う 47 問については、提案手法を用いることで逆に性能が若干低下した。“+pattern” については、悪化したもの 6 問、改善したもの 4 問であった。悪化したものの原因を調べたところ、QF 抽出の失敗 (1 問)、単位判定失敗 (4 問)、別候補の沸き出し (1 問) であった。“+sampling” については、1 問を改善し、質問「国立大学・学部昼間部の入学金は２０００年度からいくらになると決まりましたか。」について、常識的な回答「２７万７０００円」を「２０００円」より選好した。

総合では、QAC1 テストコレクションでは、MRR で +0.058、AFM で +0.062、と性能の改善が見られ、提案手法の有効性を確認できた。

QAC2 テストコレクション

次に、NTCIR QAC2 テストコレクションの Subtask 1 の 195 問を用いて評価を行った。質問文解析の QF の解析精度は、QF の存在する 131 問中、正しく抽出したのは 79 問 (60.3%) と、QAC1 にくらべて低い値であった。これは、QAC2 の質問の表現が QAC1 に比べて多様で、人手で記述した QF 解析用のパターン規則が対応できない質問文が多かったためである。

性能評価の結果を表 5 に示す。列 “BASE” が接尾表現だけを用いた場合、“+提案手法” が提案手法のパターン検索を用いた場合である。MRR の改善は、+0.022 と QAC1 に比べて低いが、これは QF の解析精度の影響と考えられる。正解 QF を与えた場合 (“+正解 QF”) では、+0.046 と MRR を改善できた。

3.4.3 失敗分析

QAC2 Subtask1 テストコレクションについて、提案手法のパターン検索を用いない場合 (“BASE”) の結果と、正解 QF を与えて提案手法を用いた場合 (“+正解 QF”) の結果とを比較したところ、回答の順位が改善した問題は 195 問中 27 問、逆に悪化した問題は 12 問であった。順位が悪化した 12 問について、失敗の分析を行った。表 6 に分析の詳細を示す。

失敗の原因は、QF を与えることで誤った回答が高順位で出力されることによる沸き出し誤りである。12 問中、正解と QF の意味関係が判定できなかったものが 9 問であった。これらについてはコーパスのサイズを増やすなど意味関係検査の再現率を上げ、正解の意味関係を見つけることで対処できる可能性がある。

次に、回答候補ごとに原因を調べた。QF を与えることで正解よりも上位に回答された誤った候補は 29 で、そのうち QF との間の意味関係が正しくないものは 15 であった。15 のうちの 11 候補は、関係判定に用いる正規表現パターンが不十分で文から誤った表現を切り出したためであった。例えば、「都道府県庁所在地の読売新聞各本支社」から「所在地の読売新聞」を見つけて、意味関係を誤って検出した。これらはパターンを精錬することで対処可能である。残りの 4 候補は、「カナダ作品」のように使用したパターンが意図した関係以外を表すもの、「仏山」(ホトケヤマ) のように意味の異なる単語が存在するもの、「夢のような宝石」のような慣用的表現、であった。全体としては、文字数の短い、特に一文字の、QF や回答候補の場合に、沸き出し誤りが生じやすい傾向が認められた。

残りの、誤った候補であるが意味関係の正しい 14 候補の沸き出しを押えるには、基本的には正解の意味関係を正し

[6] 正解のない質問も含む点に注意。

[7] 回答出力と評価は、QAC の問題設定に従った。Subtask1 は、上位 5 位まで回答し、最も高順位の回答を対象に、その順位の逆数を得点とし、その全問題での平均値 (MRR) で評価した。Subtask2 では、出力した全ての回答を対象に、F 値を得点とし、全問題での平均値 (AFM) で評価した。QAC1 では、Subtask1,2 に共通の問題セットが用いられた。

表 6: 失敗分析

正解 QF	正解	正解の関係検査	沸き出し誤り候補数 関係正解	誤り	誤り回答候補の例 (**太字**は関係検査誤り)
都市	９都市	0(失敗)	0	3	**三次**, **十年**, **２年**
宝石	レッド・ダイヤモンド	0	0	2	**貴金属**, **夢**
演目	代書屋	0	1	0	落語
山	ピクドビュール	1(成功)	0	1	**仏**
市	明石市	1	0	1	**神戸港**
目標値	６％	0	1	1	十一億二千四百五十万トン, **一年**
車	三リッターカー	0	5	0	異種動力の車, 自動車, 乗用車, 日本, ＣＮＧ
作品	カラー・オブ・ゴッド	0	2	3	鉄道員, 少年, **カナダ**, **ベネチア**, **日本**
出身	インド	1	2	0	サウジアラビア, スリランカ
ヒット曲	港町十三番地	0	3	0	暗夜航路, 涙の鎖, 木曾路の女
動物	オカピ	0	0	2	**多摩**, **天王寺**
所在地	大阪市中央区	0	0	2	**読売新聞**, **高松**

く判定する (再現率を上げる) か、評価関数 L の別の基準、例えば候補のコンテキストに関する基準、の精度を上げる必要がある。また、「車」などの文字数の少ない QF は、接尾表現の一致により「自動車」「乗用車」など、一般名詞が選ばれやすくなるという問題も認められた。

3.5 手法の位置付けと関連研究

3.5.1 手法の位置付け

提案手法の利点は、質問文に直接現れる語 (QF) を用いて高精度に意味関係検査を行うことにある。従来法の固有表現抽出を用いる手法が、あらかじめ定義・体系化しておいた意味カテゴリを用いるのに対して、提案手法は質問に応じて動的に意味カテゴリ (QF) の粒度を決定する手法である。例えば、QAC1 のテストコレクションから、「通信三社」「平均視聴率」「ＮＨＫ大河ドラマ」「力士」「レトルト食品」「官僚」「省庁」「入学金」「日銀総裁」「滑走路」などの QF が質問文から得られるが、このような粒度の小さいカテゴリを従来法の固有表現抽出で用いることは稀である。また、このような語は特定性が高いため、既存のシソーラスに含まれていることも期待できない。一方提案手法は、大規模なコーパスが持つ高い再現性を利用して、特定性の高い事例を扱う手法である。

事前知識構築が必要ないことも提案手法の利点である。例えば、従来法で新語に対応するためには、知識を更新・再構築する必要があるが、提案手法では新規コーパスを用意するだけで対応できる。提案手法は、上位下位関係の検査に限らずコーパスに現れる種々の意味関係を利用するが、この過程はコーパスに潜在する「常識」を利用していると考えることができる。「〜滑走路はどのくらいですか。」という回答の単位が明示されない質問から、「…滑走路は３５００メートル…」などの表現を見つけることで、単位や数値の常識的な値を選択する。

3.5.2 コーパスからの意味関係の獲得と検査

表層的な表現のパターンを利用してコーパスから意味的関係のある単語の組を自動獲得する手法は、文献 [6] をはじめ種々試みられている。また、抽出した知識を質問応答に利用した報告もある [3]。これらの従来法はすべて「獲得」の手法であったのに対して、提案手法は同様の手法を関係の有無の「検査」に用いている点が異なる。

一般に情報処理の問題としてとらえた場合、「獲得」とはある制約を満たすオブジェクトの組をすべて求める手続き、「検査」とはある特定の組について制約を満たすかどうか調べる手続き、と区別される。「獲得」の問題点は、計算コストおよび空間コストが高価なことにある。獲得の対象 (コーパス等) が大規模になる場合、現実的には獲得の範囲を制限する必要がある。そのため、対象が持つ情報は多少なりとも捨象されてしまう。例えば、自然言語処理では、獲得の対象を単語などの単純な単位に限定することが多い。

一方、特定の組が既知である場合、「獲得」を行わずに直接「検査」を行う方がずっと低コストである。また「獲得」と異なり情報が失われることはない。提案手法は、質問応答という問題設定では特定の組 (QF と回答候補) が既知であることを利用して、低コストで対象の持つ情報を失わない「検査」を活用した手法と考えることができる。

4 期待効用に基づく回答群の決定

4.1 問題設定

質問 q について質問応答を行って得られた、正解集合 A を含むと考えられる十分大きな回答候補集合を $C = \{c_1, c_2, \cdots c_n\}$ とする。C 中の各候補 c_i は、評価関数 $L(c_i|q)$ によって回答らしさを表すスコアが付与されている。ここで、C について、以下の二つの仮定を行う。

a. 回答の重複がない。任意の 2 つの候補 $c_i, c_j \in C (i \neq j)$ について、c_i と c_j は同一解ではない。ここで、同一解とは、テストコレクションの正解に基づいて定義されるとする。すなわち、テストコレクションによって同一の正解と判定されるような、複数の回答を含まない。

b. 全ての正解を含む。すなわち、$A \subset C$。

このとき、回答群の決定問題は、回答群として適切な C の部分集合 $C_s (\subset C)$ を求める問題と定義できる。

ここで、2 つの仮定は次の方法で近似的に成り立たせることができる。仮定 a は、初期回答候補集合を、重複解を見つけスコアの高い候補のみを残すような前処理を適用することによって成立させる。仮定 b は、$|C|$ を十分大きく取ることによって、近似的に成立させる。

4.2 期待 F 値の計算

まず、正解の数 $|A|$ が分かっている場合を考え、$|A| = m$ とする。回答集合 C_s を選択した時、正解数 $|A|$、回答数 $|C_s|$、回答中の正解数 $|A \cap C_s|$ から、(QAC や TREC で) リスト型質問の評価に用いられる F 値 $F(|A|, |C_s|, |A \cap C_s|)$ は、次のように計算できる。

$$F(|A|, |C_s|, |A \cap C_s|) = \frac{2 \cdot \frac{|A \cap C_s|}{|A|} \cdot \frac{|A \cap C_s|}{|C_s|}}{\frac{|A \cap C_s|}{|A|} + \frac{|A \cap C_s|}{|C_s|}}$$

F 値の期待値 $E(C_s \mid |A| = m)$ は次のように計算できる。

$$E(C_s \mid |A| = m) = \sum_{k=1}^{m} P(C_s, k \mid |A| = m) F(m, |C_s|, k)$$

ここで、$P(C_s, k \mid |A| = m)$ は、$|A| = m$ のとき C_s に丁度 k 個の正解が含まれる条件付き確率である。

次に、正解の数 $|A|$ がちょうど m となる事前確率 $P(|A| = m)$ を導入する。回答集合 C_s を選んだときの F 値の期待値は次の式で計算できる。

$$E(C_s) = \begin{cases} P(|A| = 0) \cdot 1 & \text{if } C_s = \{\} \\ \sum_{i \geq 1} P(|A| = i) E(C_s \mid |A| = i) & \\ & \text{otherwise} \end{cases}$$

ここで上段の $C_s = \{\}$ の場合は、「正解のない質問に対してはシステムが何も回答しなかった場合にのみ得点 1 を与える」という QAC の問題設定に従った期待値計算式となっている。

期待値 $E(C_s)$ を最大化する C_s を選ぶことで、回答候補群 $\hat{C}_s$ を決定する。

$$\hat{C}_s = \text{argmax}_{C_s \subset C} E(C_s)$$

4.3 確率分布の推定

4.3.1 $P(C_s, k \mid |A| = m)$ の推定

n 個の要素を持つ集合を考え、m 個がある属性 (ここでは正解) を持つとする。この集合から l 個を無作為に抽出したとき、ある属性をもつ要素を k 個抽出する確率 $P_{hg}(n, m, l, k)$ は、超幾何分布として知られ、次の式で示される。

$$P_{hg}(n, m, l, k) = \frac{\begin{pmatrix} m \\ k \end{pmatrix} \begin{pmatrix} n-m \\ l-k \end{pmatrix}}{\begin{pmatrix} n \\ l \end{pmatrix}}$$

超幾何分布は、選んだ要素を等しく扱い、ある属性を持つ確率は一様に等しいと仮定している。一方、C_s の各要素は、正解らしさを表すスコアが与えられているので、スコアが大きいほど正解である確率が高くなるよう、次のように超幾何分布を修正し、条件付き確率 $P(C_s, k \mid |A| = m)$ を近似的に求めた。

$$P(C_s, k \mid |A| = m) = \frac{\sum_{X \in \text{sel}(C_s, k)} \sum_{Y \in \text{sel}(C - C_s, m-k)} \delta(X \cup Y)}{\sum_{Z \in \text{sel}(C, |C_s|)} \delta(Z)}$$

ここで、$\text{sel}(X, i)$ は、集合 X から i 個の要素を選ぶ組み合わせ全体の集合、$\delta(X)$ は、回答候補集合 X に含まれる全要素の修正スコアの総和で、次の式で表される。

$$\delta(X) = \sum_{x \in X} L'(x|q)$$

修正スコア $L'(x|q)$ は、評価関数によるスコア $L(x|q)$ を、$x \geq 0$ で定義される単調非減少関数で修正したスコアである。スコア $L(x|q)$ は値の比較 (順序関係) だけを考慮したものであるため、値の大きさを考慮するように修正が必要となる。本稿では次のような修正関数を導入した。

$$L'(x|q) = L(x|q)^{\alpha}$$

4.3.2 $P(|A| = m)$ の推定

正解数の事前分布は、質問テストセットと同じ傾向を持つテストコレクションを用いて、その標本分布から推定することも可能である。しかし、質問の傾向は、それを発する人物、その人の抱える情報要求、対象とするタスク、などに大きく依存するため、コレクションの統計情報からの推定では高い精度が得られない。例えば、QAC1 と QAC2 では、正解数の分布が大きく異なっている。

本稿では、質問を解析して手がかりとする方法として、次の事前分布を用いた。

$$P_a(|A| = m) = \begin{cases} p_0 & \text{if } m = 0 \\ 1 - p_0 & \text{if } m = e \\ 0 & \text{otherwise} \end{cases}$$

ここで p_0 は、回答数が 0 となる確率である。テストコレクションの標本分布から求めるか、先験的に決める。

また e は、質問文を解析して得られる回答数の予測である。回答数の予想は、次の手法によって求めた。

- もし、QF 中に数値が含まれていたら、それを回答数の予想とする。例えば、「通信三社」から回答数 3 を予想する。
- 質問中に疑問詞が 2 つ以上存在する場合、その回数を予測とする。例えば、「〜誰と誰ですか」から回答数 2 を予想する。
- それ以外の場合、回答数を 1 と予想する。

4.4 評価実験

QAC テストコレクションを用いて、回答数を選択する次の 2 つの戦略を比較した。

BEST(n) スコアの高い順に n 個回答する。($n = 1 \cdots 5$)

UMP(α) 期待効用最大化によって回答数を決定する。

UMP のパラメータは、$p_0 = 0$ を固定し、$0 < \alpha \leq 5$ で変化させた。また回答数の上限を 5 に設定した。各戦略のにおける平均回答数と平均 F 値 (評価指標) との関係を図 4 に示す。どの平均回答数についても提案法の UMP が BEST を上回っており、問題によって適切に回答数を選択していることがわかる。

4.5 手法の性質と関連研究

4.5.1 候補群の選択

本稿で用いた条件付き確率 $P(C_s, k \mid |A| = m)$ では、C_s 中の各要素の選び方は他の要素とは独立と仮定しており、C_s に含まれる特定の要素の組合わせによって値が増減することなく、個々の要素のスコアによってのみ決まるような確率モデルを用いた。回答候補集合 $c_1, \cdots, c_n \in C$ が $L(c_i|q)$ によって降順にソートされているとすると、$E(C_s)$ を最大化する C_s は、スコアの高い候補から順に $j \leq 0$ 個選んだ $C_s = \{c_1 \cdots c_j\}$ となる。したがって、本稿の範囲では、提案手法は回答数 j を求める手法となっている。

しかしながら、提案した枠組みは、候補群を選ぶ手法となっている点に注意されたい。$P(C_s, k \mid |A| = m)$ に候補間の依存関係を用いた確率モデルを導入すると、候補の組み合わせまでを考慮にいれた回答候補選択手法となる。例えば、同じ並列句構造に現れたり、同じ意味カテゴリに属する複数の候補は共に正解となりやすい、といった基準を利用した確率モデルを導入できる。

4.5.2 効用関数

本稿で効用関数に F 値を用いたのは、QAC2 の評価指標に従ったためである。同様に、効用関数として他の評価指標を導入することも可能である。例えば、以下の重み付き F 値を効用関数として用いることで、再現率 ($\beta > 1$) や精度 ($\beta < 1$) を重視した回答候補の選択が実現できる。

$$F(|A|, |C_s|, |A \cap C_s|) = \frac{1 + \beta^2}{\frac{|A|}{|A \cap C_s|} + \frac{\beta^2 |C_s|}{|A \cap C_s|}}$$

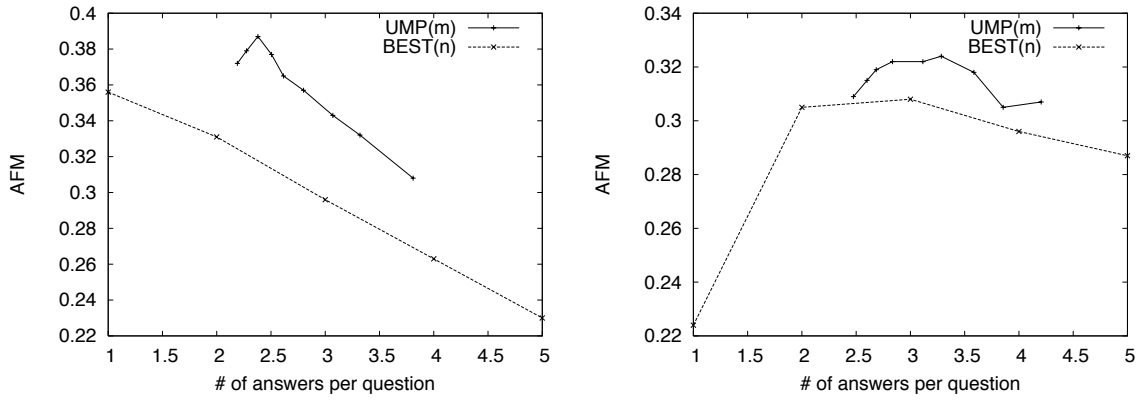

図 4: QAC1(左) と QAC2(右) に対する平均回答数と平均 F 値 (AFM) の関係

4.5.3 関連研究

リスト型質問は、NTCIR では QAC1(2002) から、TREC では TREC2003 からと、評価が始まったのが比較的最近であることもあり、まだそれほど多くの研究は行われていない。文献 [2] では、本稿の手法のように期待効用を用いた手法を提案しているが、その定式化は大きく異なる。文献 [5] では、回答間の依存関係を利用した手法を提案しているが、選択はヒューリスティクスな手法を用いている。その他、NTCIR の QAC や TREC の QA Track に参加したシステムでは、単純なヒューリスティクスを用いた手法がほとんどである。

5 むすび

質問応答の要素技術として、2 つの手法を提案した。

第 1 に、回答候補の意味的な妥当性を検査する手法として、コーパスに遍在する「常識」を利用した手法を提案した。この手法は、質問文の中心的な語 (QF) を利用し、意味カテゴリ粒度を動的に決定するため、高精度の検査が可能な手法であった。また、素のコーパスを知識源として利用するため、事前知識構築を必要としない利点もあった。評価実験によって手法の有効性を確認する一方、さらに精度を向上させるためには、QF の解析精度の向上、関係の検査に用いるパターン規則の精錬、が必要であることが示された。今後の課題として、本稿で用いた関係の他に、同値関係など回答候補選択の手がかりとなる種々の意味関係を導入することが考えられる。また、意味関係検査手法として質問応答以外の自然言語処理への応用も課題である。

第 2 に、リスト型質問に対して、期待効用最大化原理に基づく回答群選択手法を提案した。本問題に対し理論的な意味付けを与えるとともに、評価実験により性能向上も確認した。一方、本稿で用いた確率モデルは、各候補の選択を独立としているため、回答数を求める手法に止まっている。より適切な候補群選択を行うため、回答候補間の依存関係を考慮した確率モデルの導入が今後の課題である。

参考文献

[1] T. Akiba, K. Itou, and A. Fujii. Question answering using "common sense" and utility maximization principle. In *Working Notes of 4th NTCIR workshop*, pp. 297–303, 2004.

[2] J. D. Burger. MITRE's qanda at TREC-12. In *Proceedings of TREC-12*, pp. 436–440, 2003.

[3] M. Fleischman, E. Hovy, and A. Echihabi. Offline strategies for online question answering: Answering questions before they are asked. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*, pp. 1–7, 2003.

[4] J. Fukumoto, T. Kato, and F. Masui. Question answering challenge for five ranked answers and list answers – an overview of NTCIR 4 QAC2 subtask 1 and 2 –. In *Working Notes of 4th NTCIR workshop*, pp. 283–290, 2004.

[5] J. Fukumoto, T. Niwa, M. Itoigawa, and M. Matsuda. Rits-qa: List answer detection and context task with ellipses handling. In *Working Notes of 4th NTCIR workshop*, pp. 310–314, 2004.

[6] M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of International Conference on Computational Linguistics*, pp. 539–545, 1992.

[7] E. Hovy, U. Hemjakob, and C.-Y. Lin. The use of external knowledge in factoid qa. In *Proceedings of TREC-11*, pp. 644–652, 2001.

[8] H. Isozaki. NTT's question answering system for NTCIR QAC2. In *Working Notes of 4th NTCIR workshop*, pp. 326–332, 2004.

[9] A. Ittycheriah and S. Roukos. IBM's statistical question answering system – TREC-10. In *Proceedings of TREC-10*, pp. 258–264, 2001.

[10] S. Lee and G. G. Lee. SiteQ/J: A question answering system for Japanese. In *Proceedings of The third NTCIR Workshop*, 2003.

[11] M. Murata, M. Utiyama, and H. Isahara. A question-answering system using unit estimation and probabilistic near-terms IR. In *Proceedings of The third NTCIR Workshop*, 2003.

[12] J. Prager and J. Chu-Carroll. Answering what-is question by virtual annotation. In *Proceedings of Human Language Technology Conference*, pp. 26–30, 2001.

[13] T. Takahashi, K. Hawata, S. Kouda, and K. Inui. Seeking answers by structural matching and paraphrasing. In *Proceedings of The third NTCIR Workshop*, pp. 87–94, 2003.

[14] S. Tellex, B. Katz, J. Lin, A. Fernandes, and G. Marton. Quantitative evaluation of passage retrieval algorithms for question answering. In *Proceedings of ACM SIGIR*, pp. 41–47, 2003.

[15] E. Voorhees and D. Tice. The TREC-8 question answering track evaluation. In *Proceedings of the 8th Text Retrieval Conference*, pp. 83–106, Gaithersburg, Maryland, 1999.

[16] 倉田, 岡崎, 石塚. 係り受け関係に基づくグラフ構造を用いた質問応答システム. 信学技法 NLC2003-35, pp. 1–7, 2003.