質問応答における文書集合を用いた常識的な解の選択

秋葉友良 † 伊藤克亘 ‡ 藤井敦 *

† 産業技術総合研究所 ‡ 名古屋大学 * 筑波大学

e-mail: t-akiba@aist.go.jp

1 はじめに

米国 NIST の TREC-8(1999) 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。

質問応答では、関連文書から回答候補を抽出する際に、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 [7]、既存のシソーラス [10]、事前獲得した知識 [4] などを用いて、回答候補の意味的妥当性を判定する。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一目瞭然である。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文書集合中には「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、NTCIR-4 の QAC2 に参加した筆者らの質問応答システムについて、回答候補の妥当性を判定するために文書集合をそのまま知識源として用いる新規手法を提案する。提案手法は文書集合を一般知識源として用いるので、質問応答の対象以外の文書も利用可能で、かつ文書をなるべく多く使うほど効果が期待できる。

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 (1) 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。

質問応答 (2) 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 について、q の回答集合としての $A \in 2^S$ の相応しさを表す評価関数 $L(A|q)$ によって、回答集合 $\hat{A} = \mathrm{argmax}_{A \in 2^S} L(A|q)$ を求める。

質問応答 (1) は、正解を一つ見つける問題で、TREC の factoid question や NTCIR QAC の Task1 に対応する。質問応答 (2) は、複数の正解を網羅的かつ誤りなく見つける問題で、TREC の list question や NTCIR QAC の Task2,3 に対応する。実際に $\hat{a}$ や $\hat{A}$ を求めるには、巨大な空間 S を探索するため、関連文書検索および、文解析 (形態素解析等) や固有表現抽出による回答候補抽出によって、探索空間の絞り込みが行われる。

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 [12]。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 [7]。

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タイプ (あるいは意味的なカテゴリ) 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タイプ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 [7] では、独自に設定した 62 のカテゴリを利用しており、NTCIR QAC1 評価で好成績を達成した。また、文献 [11] では階層的に分類した 80 以上のカテゴリを用いている。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 [6] 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。
- 質問文や対象文書に現れる具体表現から、抽象的なカテゴリへのマッピングを行うため、具体表現の持つ情報が失われる。例えば、「１９９８年と１９９９年の２年間に横綱に昇進した力士の名前は何ですか。」という質問の場合、「横綱」や「力士」というカテゴリが設けられていない限り、よくても「スポーツ選手」などのカテゴリに抽象化され、「野球選手」などとの区別ができない。

本稿の提案手法は、b「回答候補自体に関する基準」について、固有表現抽出のカテゴリ詳細化の代わりに、文書集合と検索エンジンを用いる手法である。

3 文書集合と検索エンジンを用いた語と語の意味的関係の検査

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、「２００

０年のＮＨＫ大河ドラマは何ですか。」という質問には、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。また、質問「ＺＩＰの記憶容量はいくつですか。」という質問では、「記憶容量」という語により正解の数に後続する単位が「MB(メガバイト)」「GB(ギガバイト)」などであろうことが示唆される。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。

提案手法は、質問文解析によって抽出した QF と回答候補の、２つの文字列の間の意味的な関係の有無を直接検査して、質問応答の評価関数 L の基準として利用する。意味的関係の検査は、情報検索エンジンを利用してコーパスから特定の言語表現パターンを検索することで実現する。提案手法の特長は以下の通りである。

- コーパスをそのまま利用し、事前の知識構築を要しない。
- 高度な言語処理を用いずに実現可能。
- 抽象化されたカテゴリを介さず、質問文や回答候補に現れる表現を文字通り用いるため、高精度の意味検査が可能。

手法の実装は、まず QF と回答候補から抽出した索引語集合でコーパスを AND 検索し、抽出した文書から QF と回答候補を含むパターンを見つけることで実現した。パターンは、表層表現と形態素解析によって得られる情報 (品詞等) を用いた正規表現を人手で記述した。TREC における factoid question や、NTCIR QAC1,2 の質問では、回答として名称か数量表現を期待する。以下では、それぞれの場合の処理と用いたパターンについて説明する[1] 。

3.1 名称の検査

“ＮＨＫ大河ドラマ” と “新選組” が上位下位関係にあることは、コーパスに “ＮＨＫ大河ドラマ「新選組」” のような表現が現れることで判定できる。QF と回答候補 (AC) の間の上位下位関係の有無を、“AC という QF”、“AC 以外の QF”、“QF「AC」” といった表層表現パターンの発見によって検査する。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 [5] 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。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。

パターンを検索で発見する処理に加えて、次のような判定方法も利用する。QF と回答候補の文字列だけから上位下位関係にあることが分かる場合もある。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。このような場合を、以下の処理で検査する。

- 回答候補文字列の末尾が QF と一致するかを調べる。
- 回答候補と QF をこの順に連接した文字列が、形態素解析の辞書または検索エンジンの索引語として登録されているかを調べる。

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。

[1] 回答が名称か数量表現のどちらになるかは、質問文解析によってほぼ推定可能である。また、どちらかに決定できない場合でも、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判定できる。

単位の検査

“記憶容量” の単位として “ＭＢ” が相応しいことは、コーパスに “記憶容量は 650 ＭＢ” といった表現が現れることで判定できる。回答候補の末尾に現れる「単位」(UNIT) が、QF が示唆する単位として適切かどうかを、“QF は num $UNIT$”(num は任意の数値文字列) といった表層表現パターンを見つけることによって検査する。文献 [9] 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。

以上の処理に加えて、質問文に直接回答の単位が明示されている場合には、それも判定に用いた。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。

数値の検査

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。この過程はランダムサンプリングと見なせるので抽出される数値が正規分布に従うことを仮定し、求めた分布に対する回答候補の数値の妥当性を検査する。具体的には、サンプリングした数値集合の平均と分散から分布パラメータを決定し、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を行い、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。

4 評価実験

NTCIR QAC1 の formalrun テストコレクションを用いて、提案手法を評価した。検索エンジンには GETA[1] を使用した。意味関係を検査する文書集合は、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。

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。QAC1 formalrun Task1 の 200 問から Question Focus を人手で抽出し、正解がある質問 196 問について QF と正解を含むパターンをどの程度検出可能かを調べた。

QF と認められる文字列の候補は、一意には決まらず複数考えられることがある。特に、単独の単語から、修飾句まで含めた大きな句の単位まで、どのような単位を抽出するかという選択が考えられる。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」などの候補が考えられる。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、本実験では次の単位を QF と認定した。

- 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。
- 質問中に複数の箇所で候補が見つかった場合は、人手でより尤もらしいものを選択。

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

名称

名称を問う質問は、196 問中 149 問であった。これらについては、正解の上位語にあたるものを QF として認定した。今回は、正解と同格、同意語の関係にあるものは除外した[2] 。例えば、「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。」の「在日米軍駐留経費」は分析の対象から除外した。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。一つの質問に複数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。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%)、表層的手がかりと併用した場合 (“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 問について、正解の数量表現に現れる単位と何らかの関係を持ち、推定の手がかりとなる文字列を QF として認定した。これらには、「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「平均視聴率」のような属性を表す語に加えて、「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「滑走路」のような正解の属性を有する対象そのものを表す語も含まれる。後者の場合、「滑走路」のどの属性について訪ねているのか質問には明記されておらず、常識的な属性について回答が望まれる。33 問すべてについて QF が抽出できた。

抽出した 33 問の QF とそれぞれの正解数量表現から抽出した単位について、関係の有無を提案手法によって検査した。複数の正解がある場合は、名称の場合と同様に、一つでも関係が認められた場合に成功とした。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%)、表層的手がかりと併用した場合 (“suffix”) で 26 問 (78.8%) の関係が判定できた。

[2] 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。

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)

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。

文書集合のサイズ

提案手法は、パターン検索の対象とする文書集合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。

4.2 質問応答システムの評価

提案手法を質問応答システムに実装して、システム全体の性能を調べた。構築したシステムは次のような特徴を持つ。

- 質問文解析により、QF、単位表現、質問タイプ、検索キーワードを抽出。
- 質問タイプは、疑問詞で確実に判別できる程度の、6 種類のカテゴリに分類し、回答候補を抽出する手がかりとして利用。
- 検索エンジンには GETA[1] を用い、文書単位で検索。
- 回答候補のコンテキストに関する基準、回答候補自体の基準を組み合わせた評価関数で回答候補を選択。
- 同義の回答をまとめる手法、複数回答集合を選択する手法を用いて、list 型の質問に対応。

NTCIR QAC1 の formalrun テストコレクションを用いて評価を行った。システムの QF の抽出精度を Task1,2 の 200 問[3] 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”、“MF” は、それぞれ QAC1 の Task1(Mean Reciprocal Rank)、Task2(Mean F-measure) の結果を表す[4] 。行の “BASE” は、意味検査に提案手法を用いない場合の結果[5] 、“+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、Task1 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。

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

[3] 正解のない質問も含む点に注意。

[4] QAC1 では、Task1,2 に共通の問題セットが用いられた。

[5] 4.1 節の “suffix” のような表層的な手がかりは利用した。

表 4: 質問応答システム全体での評価

回答種別	system	MRR	MF
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

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、MRR で +0.058、MF 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。さらに性能を改善するには、QF 抽出の精度改善、検査の再現率の向上、沸き出し誤りを押さえる工夫が必要と考えられる。

5 関連研究

5.1 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 [5] 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 [4]。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[6] 。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。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。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。

5.2 Question Focus とシソーラスの利用

本稿の QF に相当する表現を質問文から抽出し、回答候補との関係を既存のシソーラスを用いて調べる手法が提案されている。文献 [8] では、質問が何を探しているか、あるいは何に関する質問なのかを表す単語 (あるいは単語列) を Question Focus と呼び、疑問詞 “what” などの回答のタイプが特定できない質問の回答抽出の手がかりとして利用している。文献 [6] では、疑問詞に後続する単語 (列) を focus として抽出し、WordNet で上位あるいは下位の関係にある語を回答候補として重視する評価基準を用いている。文献 [10] では、“What is X?” 型の質問について、X の上位語を回答候補として WordNet から抽出、X と上位語のコーパスでの共起から適切な抽象度の候補を選択する手法を示している。

5.3 大規模コーパスの利用

大規模コーパスを利用した質問応答の研究には、文献 [2, 3] がある。これらの手法は、QA の対象文書を大規模にすることの効果を利用したものである。サイズの大きな検索対象を用いた候補抽出の再現率向上と、抽出した候補の冗長性による精度向上が期待できる。抽出した候補は、最後に対象文書へと投射することで根拠となる文書を得る。これに対し提案手法は、質問応答に用いる一般知識源として大規模コーパスを利用する手法であり、上記手法とも併用が可能である。

5.4 知識源としての情報検索エンジン

本稿の手法は、検索エンジンを用いた情報検索技術を、言語処理の一般知識源として直接利用する手法と考えることができる。質問応答の利用方法として、知識処理システムの一モジュールとして用いることが考えられているが、現状では質問応答の性能向上自体が研究対象であり、未だ実現していない。本研究は、質問応答に求められるほどの高度な知識でなくても、上位下位関係の検査などのより基本的な知識源として、情報検索技術を知識処理システムの一モジュールとして利用可能であることを示したと考えられる。

6 まとめ

質問応答における回答候補の評価基準について、固有表現抽出を用いる従来法を代替する新規手法を提案した。NTCIR QAC1 テストコレクションを用いた評価実験により、性能の向上が見られ、提案手法の有効性が確認できた。

[6] 例えば、獲得の対象は単語などの単純な単位に限定する、など。

参考文献

[1] 汎用連想検索エンジン GETA. http://geta.ex.nii.ac.jp.

[2] E. Brill, J. Lin, M. Banko, S. Dumais, and A. Ng. Data-intensive question answering. In *Proceedings of TREC-10*, 2001.

[3] C.L.A.Clarke, G.V.Cormack, T.R.Lynam, G.M.Li, and G.L.McLearn. Web reinforced question answering. In *Proceedings of TREC-10*, 2001.

[4] 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.

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

[6] A. Ittycheriah and S. Roukos. IBM's statistical question answering system – TREC-10. In *Proceedings of TREC-11*, 2001.

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

[8] D. Moldovan, S. Harabagiu, M. Pasca, R. Mihalcea, R. Goodrum, R. Girju, and V. Rus. LASSO: A tool for surfing the answer net. In *Proceedings of TREC-8*, pp. 65–73, 1999.

[9] 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.

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

[11] Y. Sasaki, H. Isozaki, T. Hirao, K. Kokuryou, and E. Maeda. NTT's QA systems for NTCIR QAC-1. In *Proceedings of The third NTCIR Workshop*.

[12] 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.