

新聞記事に対する Why 型質問の回答抽出手法の提案

小林竜大^{†1} 小島俊輔^{†2}

近年、ユーザから質問を受け付け、質問に対する回答を自動抽出する質問応答システムが注目されている。しかし、Why 型質問の回答を自動抽出する有効な手法は提案されていない。従来の Why 型質問の回答抽出手法は回答が複数文続く場合抽出できないという問題がある。また、Web からの回答抽出のみを想定している手法が多い。本論文では、従来の Web からの抽出を想定した Why 型質問の回答抽出に開発された回答位置特定アルゴリズムを改良した、新聞記事に適用した回答位置特定アルゴリズムによる回答抽出手法を提案する。新聞記事データベースより評価セットを作成し、提案手法の評価実験を行った結果、適合率 53.8%、再現率 64.1% で回答を抽出した。

Proposal of Answer Extraction Method of Why Type Question for Newspaper Article

Kobayashi, Ryuta^{†1} Oshima, Shunsuke^{†2}

Recently, Question-Answering system that extracts answers of Why type questions that imputed by users is gathering attention. However, there no effective method that extracts answers of Why type questions. Conventional answer extraction method of Why type question doesn't extract multiline answer and extract answers from only Web. This paper proposes an answer extraction method of Why type question using an algorithm for finding an answer in a newspaper article, improving an algorithm for finding an answer in a Web page. As a result of evaluation experiment for the proposed algorithm using evaluation set made from newspaper article database, the proposed algorithm has 53.8% precision and 64.1% recall.

1. はじめに

近年、新聞記事などの膨大な文書の中から必要な情報を取り出すために、検索システムを利用する機会が多くなった。しかし、現在の検索システムでは、必要な情報に関連するキーワードを用いて検索を行い、検索結果を 1 件ずつみていき、ユーザが必要とする情報が存在するか調べる必要がある。これはユーザにとって大きな負担となる。そこで、ユーザから質問を受け付け、その質問に対する回答を自動抽出する質問応答システムが注目されている。

質問応答システムには、情報探索、情報抽出、自然言語処理などの様々な技術を組合せ、質問に対して適切な回答を抽出する質問応答技術が用いられている。現在、この質問応答技術について活発な研究が行われている。しかし、従来の質問応答技術では、人物 (Who)、いつ (When)、何 (What)、場所 (Where) を回答の対象とする質問に対して有効であるが、Why 型質問を回答の対象とする質問に対して有効な手段は確立されていない。質問応答技術の精度を評価するワークショップとして、アメリカ国立標準技術研究所とアメリカ国防総省内の研究部門の一つである ARDA が提供する TREC や国立情報科学研究所が提供する NTCIR の QAC がある。これらのワークショップでも、

人物 (Who)、いつ (When) などを回答の対象とする質問を多く取り扱われているが、なぜ (Why) を回答の対象とする質問は少ない。

また、質問応答システムでは回答を抽出するために、質問に関連した知識文を Web 文書や新聞記事から検索する。Web 文書は、新聞記事に比べ膨大な文書量が存在するため、Why 型質問に対する回答が存在する可能性が高いと考えられる。また、新聞記事は、主にある事実とその背景が書かれているため、正確性と信頼性が高い。このように Web 文書と新聞記事にはそれぞれ特徴がある。したがって、Web 文書と新聞記事の両方を検索源とすることで、Why 型質問に対する回答を抽出できる可能性が広がると考えられる。

本論文では、従来 Web からの抽出を想定した Why 型質問に対する回答抽出手法を新聞記事に適用させる。そして、新聞記事から Why 型質問に対する回答を自動抽出する手法を提案する。

2. 従来の研究とその問題点

本章では、Why 型質問に関する従来の研究について説明し、その問題点を示す。

2.1 Why 型質問に関する従来の研究

田村らの研究¹⁾では、質問応答システムに入力された Why 型質問文からキーワードと回答抽出パターンを作成し、キーワードを検索エンジンに与えることで概要文を得て、概要文中で、回答抽出パターンとのマッチングした文

^{†1} 熊本高等専門学校 生産システム工学専攻
Kumamoto National College of Technology, Production Systems Engineering Course
^{†2} 熊本高等専門学校八代キャンパス ICT 活動支援センター
Kumamoto National College of Technology, ICT Center for Learning Support

を回答文として抽出している。しかし、質問文から作成されたキーワードと回答抽出パターンが長くなることで、検索結果の減少や概要文と回答抽出パターンとのマッチングの失敗が発生し、Why 型質問に対する回答を抽出できないことがある。このため、質問文によって Why 型質問に対する回答候補が少ない場合がある。

諸岡らの研究²⁾では、質問応答システムに入力された Why 型質問からキーワードを抽出し、そのキーワードを検索エンジンに与えることで概要文を得て、さらに、概要文中の、「動詞＋ため」などの抽出パターンとマッチングする、かつ「動詞＋ため＋助詞(の)」などの非抽出パターンとマッチングしない文を回答文として抽出している。しかし、回答抽出の対象となる文が限定され、例えば、文章中の Why 型質問に対する回答を表す文が複数文続く場合、全ての回答を抽出することができない。

渋沢らの研究³⁾では、質問応答システムに Why 質問の内容を最も表す検索キーワードを入力し、文章中で検索キーワードを全て含む文を Why 型質問に相当する文とし、Why 型質問の内容を表す文と回答文の相対的な位置関係を特定することで、Why 型質問に対する回答を抽出している。回答文との相対的な位置関係は Why 型質問に相当する文に出現する特徴語の組合せによって特定される。この特徴語の組合せによって回答文との相対的な位置関係を特定し、回答を抽出する手法を回答位置特定アルゴリズムという。渋沢らが提案している特徴語の種類と例を表 1 に示す。また、相対的な位置関係を特定できる特徴語の組合せを表 2 に示す。

また、田村らの手法と渋沢らの手法は、Web 文書を検索源としており、諸岡の手法は、新聞記事を検索源としている。

田村らの手法と諸岡らの手法は、回答抽出の対象となる文が限定され、回答が複数文続く場合には対応できないと

表 1 特徴語の例

特徴語の種類	特徴語の例
理由語	理由、原因、ため、用言基本形＋から
前方指示語	それ、これ、以上、その、こんな
後方指示語	以下、次、以後、後述、示す

表 2 回答位置を特定できる特徴語の組合せ
(理：理由語、前：前方指示語、後：後方指示語)

理	前	後	回答の位置
○	○	×	質問に相当する文の前方
○	×	○	質問に相当する文の後方
×	×	×	質問に相当する文の後方 (後方に理由語がある場合のみ)
○	○	×	質問に相当する文と同一

いう問題がある。渋沢らの手法は、Why 型質問の内容を表す文と回答文の相対的な位置に注目することで回答が複数文続く場合でも抽出可能になる。しかし、この手法は Web 文書からの検索を想定しており、独特の言い回しが含まれた新聞記事から回答位置特定アルゴリズムにより回答を抽出すると精度が低くなる。以下に、回答位置特定アルゴリズムの新聞記事への適用における問題点を述べる。

2.2 新聞記事への適用における問題点

新聞記事は独特の言い回しで記述されている。その言い回しには、ある事実とその背景を端的に述べた文章、人物の談話や手紙の内容などの引用を含む文章、インタビューや質疑応答の記録文などがある。その例を図 1 に示す。例 1 と例 3 では特徴語が存在しないため抽出することができず、例 2 では特徴語が後方指示語のみなので、回答の位置を特定できない。このように、独特の言い回しを含む文では、特徴語が存在しない、または回答の位置を特定できる特徴語の組合せが存在しないことがある。

また、新聞記事には新聞記事特有の理由語が存在する。図 2 に例を示す。例では、背景を理由語として登録していないため、回答位置を特定できない。

3. 提案手法

2.2 節で述べた問題点を解決するために、新たな特徴語を登録することで回答位置特定アルゴリズムを拡張した。

例 1：ある事実とその背景を端的に述べた文章

優勝の価値が薄くなると、多くのサポーターは J リーグの 2 ステージ制に反対しています。

例 2：人物の談話や手紙の内容などの引用を含む文章

鈴木太郎選手の「引退の秘話」全文は次の通り。
「モチベーションが低下し、ただお金をもらいプレイしているだけだと感じた・・・(省略)」

例 3：インタビューや質疑応答の記録文

「なぜ、消費税増税を実施するのか。」
「所得税や法人税の引き上げは、現役世代に負担が集中し不景気では税収が少なくなる。消費税は全世代で負担でき、不景気での影響も小さく、高齢化社会における社会保障の財源にふさわしい。」

太字：Why 型質問に相当する文 二重下線：後方指示語

図 1 新聞記事の独特の言い回しの例

アップルの利益が 4 四半期連続で前年割れとなった背景
に、低価格のスマートフォンを売り出す中国勢との激しい競争がある。

太字：Why 型質問に相当する文 四角囲：新聞記事特有の理由語

図 2 新聞記事特有の理由語が出現する例

この拡張された回答位置特定アルゴリズムを用いて新聞記事から Why 型質問に対する回答を抽出する手法を提案する。

3.1 新たな特徴語の登録

本研究の提案手法では、回答位置特定アルゴリズムに以下の4つの特徴語を登録し、Why 型質問に相当する文と回答文の位置関係を特定できるようにした。

- (1) 読点を含む理由語
- (2) 引用開始語
- (3) 質疑応答開始語
- (4) 新聞記事特有の理由語

以下に、各特徴語の詳細について述べる。また、(1)～(3)では特徴語が出現した場合の Why 型質問に相当する文と回答文の位置関係についても述べる。

(1) 読点を含む理由語

「動詞（基本形）＋接続助詞（と）＋読点（,）」を、読点を含む理由語として登録した。「動詞（基本形）＋接続助詞（と）」は、「すると」のような順接の接続詞を表す。

「すると」のような順接の接続詞の後に読点が打たれており、読点の後方の文が Why 型質問に相当する文である場合、接続助詞（と）の前方の文は Why 型質問に関連した内容である可能性が高い。したがって、文中で「ある文＋動詞（基本形）＋接続助詞（と）＋読点（,）＋Why 型質問に相当する文」とマッチングしたとき、回答文は Why 型質問に相当する文と同一であると特定する。

(2) 引用開始語

新聞記事では、ある人物の談話や手紙の内容などを引用するとき、「（ある人物の談話は）次の通り。」のように引用開始を示す文を記述して引用を記載する。そのため、「次の通り。」を引用開始語として登録した。

引用開始を示す文に Why 型質問に相当する文が含まれている場合、引用された文は Why 型質問に関連した内容である可能性が高い。したがって、Why 型質問に相当する文中で、「次の通り。」が出現したとき、Why 型質問に相当する文の後方にある引用された文を回答文として特定する。

(3) 質疑応答開始語

新聞記事では、インタビューや質疑応答の記録文を記載する場合、質問文は鍵括弧（「」）が付けられている。そのため、鍵括弧（「」）を質疑応答開始語として登録した。

質問文に Why 型質問に相当する文が含まれている場合、その質問文に対する応答文は Why 型質問に関連した内容である可能性が高い。したがって、Why 型質問に相当する文の文頭と文末に鍵括弧が出現したとき、Why 型質問に相当する文の後方にある応答文を回答文として特定する。

(1)～(3)の特徴語が出現した場合の Why 型質問に相当する文と回答文の位置関係を表3に示す。

(4) 新聞記事特有の理由語

新聞記事特有の理由語について調べた結果を表4に示す。

表3：新たに登録した特徴語による回答位置の特定
(読：読点を含む理由語，引：引用開始語，質：質疑応答開始語)

読	引	質	回答の位置
○	-	-	質問に相当する文と同一
-	○	-	質問に相当する文の後方
-	-	○	質問に相当する文の後方

表4 新たな理由語の一覧

理由語の種類	理由語の一覧
新聞記事特有の理由語	背景，核心，引き金，副作用，要因，助動詞(基本形)＋から

表4を理由語として登録し、回答位置特定アルゴリズムによる回答抽出の精度を上げる。

3.2 処理の流れ

提案手法では、ユーザによる入力 Why 型の質問文ではなく、Why 型質問の特徴を最もよく表すキーワードとする。このキーワードを検索キーワードと呼ぶ。また、検索キーワードを全て含む文章を質問相当文と呼ぶ。提案手法の処理は、新聞記事に対して(1)テキストセグメンテーションを実行し、テキストセグメンテーションを行った文章に対して(2)回答位置特定アルゴリズムにより回答を抽出するという流れで行う。以下に、テキストセグメンテーションと回答位置特定アルゴリズムの各処理について述べる。

(1) テキストセグメンテーション

テキストセグメンテーションとは、文章を意味的なまとまりごとに分割する手法である。この文章の意味的なまとまりをテキストセグメントと呼ぶ。

提案手法では、新聞記事に対してテキストセグメンテーションを行い、新聞記事をテキストセグメントに分割する。そのテキストセグメントに回答位置特定アルゴリズムによる処理を行い、質問相当文と回答文の位置関係を特定する。ここで、質問相当文の前後に回答文があると特定された場合、テキストセグメントの境界までを回答文として抽出することで、質問相当文と意味的に関係のある文を抽出することができる。

テキストセグメンテーションの方法は、平尾らの研究⁴⁾などで提案されている。本論文では新たにテキストセグメンテーション法を提案しない。しかし、従来のテキストセグメンテーション法は精度が低いため、別の研究テーマとして追求する必要がある。

(2) 回答位置特定アルゴリズム

テキストセグメント中に質問相当文が存在する場合、質問相当文に出現する特徴語を調べ、表2および表3とマッチングすることで質問相当文と回答文の位置関係を特定する。ここで、質問相当文に出現する特徴語が表2および表3に同時にマッチングした場合、表3を優先して位置関係

を特定する。これは、表 3 は新聞記事の独特の言い回しに対応できるよう登録した特徴語であるため、質問相当文と回答文の位置関係を正しく特定できる可能性が高いからである。回答文が質問相当文の前方もしくは後方にあると特定された場合、テキストセグメントの境界までを回答文として抽出する。また、質問相当文と回答文が同一であると特定された場合、質問相当文を回答文とみなして抽出する。

4. 評価実験

新聞記事に適用した回答位置特定アルゴリズムと従来の Web ページを想定した回答位置特定アルゴリズムによる回答抽出の精度を比較する評価実験を行った。以下に評価セットと実験方法について説明する。

4.1 評価セット

評価セットは熊本日日新聞記事データベース⁵⁾から人手で作成した。評価セットは質問相当文を含むテキストセグメントの集合で構成される。さらに、評価セットは回答文を含む正解セットと、回答文を含まない、または質問相当文が Why 型質問の内容を表していない不正解セットに分類される。

評価セット作成のために、A.1 に示す 18 問の Why 型質問を用意した。18 問の質問は、熊本日日新聞の新聞記事が分類される 9 つの記事ジャンルから、各ジャンルに関する質問を 2 問ずつ用意したものである。A.1 に示す 18 問の各 Why 型質問に人手で検索キーワードを設定した。この検索キーワードを用いて熊本日日新聞記事データベースから記事検索を行った。記事検索すると、検索キーワードを全て含む新聞記事が新着順に出力される。記事検索の結果、検索キーワードによって膨大な量の新聞記事が出力されることがある。膨大な量の全ての新聞記事を用いて人手で評価セットを作成すると、人と時間のコストがかかる。そのため、各 Why 型質問に対する評価セットは、記事検索の結果から手動でランダムにテキストセグメントを抽出し、合計 10 個のテキストセグメントとなるよう構成した。

評価実験では、従来の回答位置特定アルゴリズムと新聞記事に適用した回答位置特定アルゴリズムの精度を評価するため、テキストセグメンテーションは全て成功していると想定する。そこで、新聞記事の形式段落を 1 つのテキストセグメントとして抽出し、評価セットを作成した。これは、新聞記事の形式段落は記事における意味的なまとまりが明確であるため、形式段落とテキストセグメントが一致すると考えるからである。

以上のように評価セットを作成した結果、評価セットは 18 問の Why 型質問について質問相当文を含む 180 個のテキストセグメントとなった。また、180 個のテキストセグメントのうち正解セットは 78 個、不正解セットは 102 個となった。

4.2 実験方法

評価セットの各テキストセグメントに対して、従来の回答位置特定アルゴリズムと新聞記事に適応した回答位置特定アルゴリズムを用いて回答を抽出する。回答位置特定アルゴリズムでは、テキストセグメントを形態素解析することで形態素列へと変換し、各特徴語とのマッチングすることで質問相当文と回答文の位置関係を特定する。形態素解析では形態素解析ツール MeCab⁶⁾を用いた。

抽出された回答の評価は人手で行う。評価基準として、出力された回答が正解を表している場合、出力された回答が正解を表しているが、質問に関係のない内容を含む場合、出力された回答が正解の一部を含んでおり、その一部から正解を推測できる場合を正解とする。

4.3 評価指標

検索性能分野の一般的な評価指標である適合率と再現率を用いる。適合率(Precision)、再現率(Recall)は以下の式で求められる。

$$\text{適合率} = \frac{\text{出力された正解数}}{\text{全出力数}} \quad (1)$$

$$\text{適合率} = \frac{\text{出力された正解数}}{\text{正解セット}} \quad (2)$$

4.4 実験結果

従来の手法により出力された回答は 71 個、そのうち、正解セットに含まれるものは 32 個となった。以上より、従来の回答位置特定アルゴリズムでは適合率 45.1%、再現率 41.0%となった。

提案手法により出力された回答は 93 個、そのうち、正解セットに含まれるものは 50 個となった。以上より、提案手法の適合率における適合率 53.8%、再現率 64.1%となった。

提案手法により正しく正解または不正解と判定されたものは、それぞれ 50 個と 62 個となった。したがって、評価セット 180 個のテキストセグメントのうち、112 個が正しく判定され、68 個が誤って判定された。また、正解セット 78 個のテキストセグメントのうち、28 個が誤って判定され、不正解セット 102 個のテキストセグメントのうち、40 個が誤って判定された。

また、提案手法で新たに登録した特徴語により 26 個の回答が抽出された。そのうち、18 個が正解であった。各特徴語により抽出された回答数とその正解数を表 5 に示す。

表 5 各特徴語によって抽出された回答抽出数とその正解数

特徴語の種類	回答抽出数(正解数)
読点を含む理由語	2 (2)
引用開始語	2 (2)
質疑応答開始語	8 (5)
新たに登録した理由語	14 (9)
総計	26 (18)

5. 考察

評価実験の結果から提案手法の有用性と誤判定の原因について考察する。

5.1 提案手法の有効性

提案手法は、従来の手法と比較して、適合率では 8.7%、再現率では 23.1%、回答抽出の精度が向上しており、提案手法の有効性がわかる。

表 5 より、読点を含む理由語と引用開始語により抽出された回答は全て正解であった。また、質疑応答開始語により抽出された回答は 8 個、そのうち 5 個は正解であった。新たに登録した理由語により抽出された回答は 14 個、そのうち 9 個は正解であった。以上より、新たに登録した特徴語による回答抽出は、精度が高いことがわかった。

表 2 の特徴語による回答抽出より、表 3 の特徴語による回答抽出を優先した結果、4 個の回答が優先して抽出された。そのうち 2 個は、表 2 の特徴語による回答抽出では誤判定であったが、表 3 の特徴語による回答抽出では正解であった。また、その他の 2 個は、表 2 の特徴語による回答抽出と表 3 の特徴語による回答抽出のどちらの場合でも誤判定であった。以上より、表 2 の特徴語による回答抽出より、表 3 の特徴語による回答抽出を優先すると、回答抽出の精度が高くなることがわかった。

以下に、登録した特徴語による誤判定の原因を述べる。

5.2 新たに登録した特徴語による誤判定の原因

質疑応答開始語による回答抽出が誤判定であった場合と新たに登録した理由語による回答抽出が誤判定であった場合の例を図 3 に示す。図 3 は評価セットから引用したものである。

例 1：質疑応答開始語による回答抽出が誤判定であった場合
(質問 9)「今のくまモン 人気をどう受け止めていますか。」

「県から「新幹線元年事業」のアドバイザーになってほしいと頼まれ、引き受けた。くまモンは、全線開業を盛り上げる「くまもとサプライズ」のトレードマークのおまけとしてつくったもの。ここまで人気が出るとは予想していなかった。」

例 2：質疑応答開始語による回答抽出が誤判定であった場合
(質問 16)「日本人はなぜ長寿だと思ふか。」「あまり分析をしていないのでよく分からないが、いつまで長寿国でいられるかとの懸念も抱く。」

例 3：登録した理由語による回答抽出が誤判定であった場合
(質問 9) 県国際課は「くまモン 人気を背景に『熊本』の認知度を上げ、最終的に県産品の販売や熊本への誘客につなげたい」としている。

太字：質問相当語 二重下線：検索キーワード 四角囲：理由語

図 3 新たに登録した特徴語による誤判定の例

図 3 の例 1 では、「今のくまモン人気をどう受け止めていますか。」を質問相当文としているが、質問相当文は Why 型質問の「なぜくまモンは人気なのか。」に相当していない。そのため、質問相当文に対する回答文も正解ではない。このように、質問相当文が Why 質問の内容を表していないという問題がある。また、例 2 では、「日本人はなぜ長寿だと思ふか。」という質問相当文に対して、「あまり分析をしていないのでよく分からないが、・・・(省略)」という回答文を抽出している。しかし、回答文が質問に対して適切に応答をしていないため、回答文は正解ではない。このように、質問相当文と回答文の内容に関連性のない場合でも回答文を抽出してしまう問題がある。これらの問題を解決するには、構文解析や意味解析など高度な自然言語処理が必要であると考えられる。

図 3 の例 3 では、例 2 と同様に質問相当文が Why 型質問の内容を表していないが、理由語として登録した「背景」が出現するため、回答文として抽出してしまう。この問題を解決するためには、例えば「(質問相当文)の背景は～だ」のように主部、述部など文中の役割を考慮した理由語を登録することで、誤判定をある程度防止できると考える。

また、「背景」などの新たな理由語を登録すると回答文の抽出精度が低くなる可能性もある。そのため、理由語を登録する場合は十分検討する必要がある。

5.3 提案手法の誤判定の原因

正解セット 78 個における誤判定 28 個の原因を調べたところ、以下の 2 つにパターンに分類された。()内はパターンの個数である。

- (1) 質問相当文中の理由語により抽出が終了する (1)
- (2) 文章中に特徴語が存在しない (27)

また、不正解セット 102 個における誤判定 40 個の原因を調べたところ、以下の 2 つのパターンに分類された。

- (3) 質問相当文が質問の内容を表していない (24)
- (4) 回答文と質問相当文の内容に関連性がない (16)

新たに登録した特徴語による誤判定の原因も(3)と(4)のいずれかのパターンに属する。

以降より、各パターンについて考察する。また、正解セットにおける誤判定の例を図 4 に、不正解セットにおける誤判定の例を図 5 に示す。図 4、図 5 は評価セットから引用したものである。

(1) 質問相当文中の理由語によって抽出が終了する

回答位置特定アルゴリズムにおいて、回答文と質問相当文が同一であると判定された場合、その時点で回答を抽出する。そのため、回答文が質問相当文の前後に存在する場合、回答文を抽出することができない。

図 4 の例 1 では、実際には質問相当文の後方に回答文が存在するが、理由語として登録している「方針」が出現するため、回答文が質問相当文と同一であると特定され、回

例 1: 事実文中の理由語によって抽出が終了する例

(質問 4) 推進協議会では、特殊法人与公務員の天下りの問題を論議する方針を決定している。これまでの論議では、… (省略)

例 2: 文章中に特徴語となりうるものが存在しない例

(質問 14) 原田助教授は水俣病 発生 のメカニズムを(1)労働者の無機水銀(2)環境に放出された無機水銀の有機化(3)メチル化した水銀の魚などへの蓄積(4)人体への蓄積(5)水俣病の発生一 の五段階に分ける。

例 3: 文章中に特徴語となりうるものが存在しない例

(質問 7) オバマ米政権は、アサド政権がダマスカス郊外で 8 月 21 日に化学兵器攻撃を実行したと主張。一時は米国によるシリア攻撃が差し迫ったとみられていた。

例 4: 文章中に特徴語となりうるものが存在しない例

(質問 15) 門脇部長は「欧米人は少し太ったくらいでは糖尿病 になりにくい、日本人は小太りになるだけで糖尿病になりやすい」と話す。日本人の膵臓(すいぞう)が分泌するインスリンの量は、欧米人の半分。欧米人は肉食の歴史の中で膵臓の細胞が、インスリンを大量に分泌するようになった。

太字: 質問相当語 二重下線: 検索キーワード 四角囲: 理由語

図 4 正解セットにおける誤判定の例

例 1: 質問相当文が質問の内容を表していない例

日野原さんは米国の白人が長寿である理由の中で、日本人に比べ、塩分摂取量が極端に少なく、胃がんが少ない点を第一に挙げた。

例 2: 回答文が質問相当文の内容と関係ない例

日銀が大規模な金融緩和を決めた 4 月 4 日以降、狙いとは逆に長期金利は上昇傾向が続いていました。6 月 11 日の日銀金融政策決定会合で金利上昇を抑える具体策を打ち出さなかったため、市場は失望しています。政府の成長戦略も期待外れと受け止められ、株価を押し下げました。

太字: 質問相当語 二重下線: 検索キーワード 四角囲: 理由語

図 5 不正解セットにおける誤判定の例

答を抽出することができない。この問題を解決するためには、「方針」を「(質問相当文)は～(の|する)方針だ」のように主部、述部などの文中の役割を考慮した理由語として登録する必要があると考えられる。

(2) 文章中に特徴語が存在しない

回答位置特定アルゴリズムは、特徴語が出現しない場合、回答位置を特定することができず、回答文を抽出することができない。

図 4 の例 1 と例 2 では、質問相当文中に特徴語が出現し

ないため、回答位置を特定できない。また、例 3 では質問相当文の後方の文に理由語が存在しないため、回答位置を特定できない。これらの問題を解決するために、例 1 では、「(質問相当文)のメカニズムは～だ」などの主部や述部などの文中の役割を考慮した理由語の登録、例 2 や例 3 では、新聞記事の言い回しを考慮した理由語の登録や構文解析、意味解析など高度な自然言語処理が必要であると考えられる。

(3) 質問相当文が質問の内容を表していない

質問相当文が Why 型質問の内容を表していなくても、文が存在すれば正解か否かに関係なく、回答を抽出してしまう。

図 5 の例 1 では、日本人が長寿である理由について質問していることに對し、質問相当文が米国の白人が長寿であるという内容を表している。質問相当文が質問の内容を表していないが、質問相当文に理由語が出現するため、回答として抽出してしまう。この問題を解決するには、構文解析や意味解析など高度な自然言語処理が必要である。

(4) 回答文と質問相当文の内容に関連性がない

回答文が質問相当文の内容と関連性がない場合でも抽出されることがある。

図 5 の例 2 では、質問相当文は「長期金利は上昇傾向が続いていました。」という内容を表している。しかし、抽出された「6 月 11 日の…(省略)」という回答文は質問相当文と意味的な関連性はない。この問題を解決するためにも、構文解析や意味解析など高度な自然言語処理が必要である。

6. おわりに

本論文では、Why 型質問を対象に、新聞記事に適用した回答位置特定アルゴリズムを提案した。従来の手法の新聞記事への適応における問題点を解決するために、新聞記事の特徴に注目した。新聞記事データベースから作成した評価セットを用いて評価実験を行い、従来の手法と提案手法の適合率・再現率を求めた。従来の手法は適合率 45.1%、再現率 41.0%、提案手法は適合率 53.8%、再現率 64.1%となった。提案手法の精度は従来の手法より適合率では 8.7%、再現率では 23.1%向上し、提案手法の有用性を確認した。

回答位置特定アルゴリズムは、文章の表層的な情報のみを利用して回答を抽出しており、文章の構文や意味的内容を考慮していないため、誤判定が多くなる。この問題を解決するために、今後は主部と述部を考慮して特徴語を再編し、構文解析や意味解析など高度な自然言語処理を取り入れる予定である。

参考文献

1) 田村元秀, 村上仁一, 徳久雅人, 池原 悟: Web 検索エンジンを用いた Why 型質問応答システムに関する研究, 情報処理学会研究報告, No.4, pp15-21 (2008)

- 2) 諸岡 心, 福本淳一: Why型質問応答のための回答選択手法, 電気情報通信学会技術研究報告, Vol.105, No.594, pp.7-12 (2006)
- 3) 渋谷 潮, 林 貴宏, 尾内 理紀夫: Why型質問の回答文をWEBから自動抽出するシステムの開発と評価, 情報処理学会論文誌, Vol.48, No.3, pp.1512-1523 (2007)
- 4) 平尾 務, 北内 啓, 木谷 勉: 語彙的結束性と単語重要度に基づくテキストセグメンテーション, 情報処理学会論文誌, Vol.41, No.SIG-3, pp.24-36 (2000)
- 5) 熊本日日新聞社, 熊本日日新聞データベース, 入手先 (<https://kumanichi.com/yuryou/>), (参照 2014-02-04)
- 6) MeCab: Yet Another Part-of-Speech and Morphological Analyzer <http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html>

付録A.1 評価セット作成に用いた9つのジャンルに関連するWhy型質問18問と検索キーワード一覧

○節で述べた提案手法の評価セット作成のために用いた9つのジャンルとそれに関連する18問のWhy型質問文を示す。質問文の括弧内は検索キーワードを表している。

社会

- (1) なぜTPPに反対するのか?(TPP, 反対)
- (2) なぜ少子化が進むのか?(少子化, 進む)

政治

- (3) なぜ消費税を増税するのか?(消費税, 増税)
- (4) なぜ天下りが問題なのか?(天下り, 問題)

経済

- (5) なぜ日銀は国債を買うのか?(日銀, 国債)
- (6) なぜ長期金利が上昇しているのか?(長期金利, 上昇)

国際

- (7) なぜ米国はシリアを攻撃するのか?(米国, シリア, 攻撃)
- (8) なぜ北方領土は返還されないのか?(北方領土, 返還)

暮らし・話題

- (9) なぜくまモンは人気なのか?(くまモン, 人気)
- (10) なぜ禁煙できないのか?(禁煙, できない)

文化・芸能

- (11) なぜ富士山は世界遺産になったのか?(富士山, 世界遺産)
- (12) なぜ宮崎駿は引退したのか?(宮崎駿, 引退)

科学・環境

- (13) なぜ赤潮が発生するのか?(赤潮, 発生)
- (14) なぜ水俣病が発生したのか?(水俣病, 発生)

健康・医学

- (15) なぜ日本人は糖尿病になりやすいのか?(日本人, 糖尿病)
- (16) なぜ日本人は長寿なのか?(日本人, 長寿)

スポーツ

- (17) なぜボールは曲がるのか?(ボール, 曲がる)
- (18) なぜイチローはヤンキースに移籍したのか?(イチロー, ヤンキース, 移籍)