QA システムにおける質問数の推定と質問タイプの同定

望月　裕介　　八木　淳紀　　韓　東力
日本大学文理学部　情報システム解析学科

1.　はじめに

質問応答（QA）システムを評価するコンテストとして NTCIR[1]が有名である。今までのNRCIRのQACタスクで使用されていた質問文データについて調査した結果、１文からなる質問文に複数の質問箇所がほとんど含まれていないことが分かった。また、質問文そのものについて行われた研究[1]でも、一つの質問文に基本的には質問が 1 箇所に限定している。すなわち、今までの質問応答システムは「小沢征爾さんはいつ、どこで生まれましたか？」のような 1 文に異なる質問タイプの質問が２つ以上ある場合はどちらかの質問タイプにしか対応できず、質問文を質問数だけ文を入力しなければならなかった。

本研究では、１文に質問タイプの異なる複数の質問がある場合に、全部の質問に対応するために質問文の質問数の推定を行う。また、このような質問文の質問タイプを同定する方法を提案する。

2.　質問数の推定

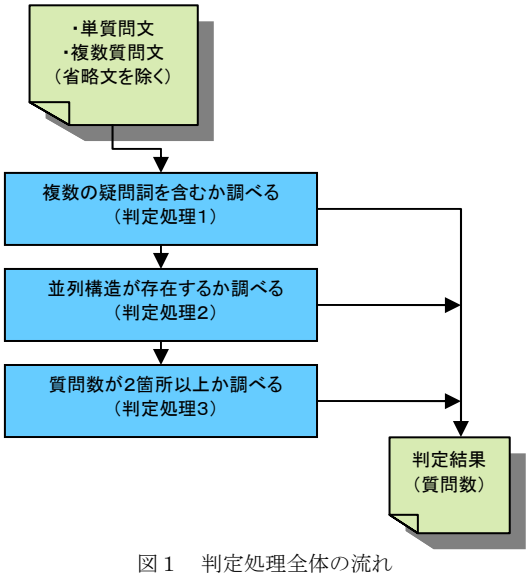

図１　判定処理全体の流れ

本研究の対象と成る質問文は、「複数質問文」と「単質問文」の２通りに分類される。複数質問文とは、例えば、「ネットオークションのメリットとデメリットは何ですか？」のような質問文の場合、「メリット」と「デメリット」のような質問箇所が２箇所以上存在する質問文のことを呼ぶ。それに対し、単質問文とは、「中南米と日本の間にはどのような関係があるのですか？」のような質問文の場合、「どのような」のような質問箇所が１箇所のみの質問文のことを呼ぶ。２章では、この複数質問文と単質問文をどのようにして判定するかを述べる。入力の対象となる質問文は複数質問文と単質問文からなる（省略文は除く）。質問文の分類は、判定処理１から判定処理３までのプロセスを通して行う。

判定処理１では、「疑問詞表現」を利用して抽出可能な複数質問文を取り除く。残りの質問文は判定処理２に渡される。判定処理２では、質問文に対して構文解析を行い、その質問文の構造を利用して抽出可能な単質問文を取り除く。残りの質問文は判定処理３に渡される。判定処理３では、Support Vector Machine（SVM）を利用して複数質問文と単質問に分類する。その際に、本研究では SVMlight[2]を使用する。最短で判定処理１のみ、最長で判定処理３までの判定処理を通過して分類される。図１は判定処理全体の流れを示したものである。

2.1.　判定処理１

質問文の形態素列に疑問詞表現が２箇所以上存在するかを調べる。この条件が真になる質問文は複数質問文になり、この条件が偽になる質問文は処理２に渡される。例えば、「現在のフランスの大統領は何代目の誰ですか。」のような質問文では、「何」と「誰」が疑問詞表現であるため、複数質問文となり、質問数は「２」となる。それに対して、「ネットオークションのメリットとデメリットは何ですか？」のような質問文では、「何」のみが疑問詞表現となり、複数質問文とは断定できないため、判定処理２に渡される。判定処理２で使用される疑問詞表現は「いつ」、「どのくらい」、「いくつ」等、全部で３３個ある。

2.2.　判定処理２

質問文に対して構文解析を行う。構文解析の結果、並列構造が存在しない場合は単質問文になり、質問の数は１と判定される。並列構造が存在する場合は処理３に渡される。一見、並列構造があれば複数質問文になり、なければ単質問文になるように思われるが、並列構造が存在しても単質問文になる質問文が存在する。それは、「予想最高気温や最低気温はどうやって予想するのですか？」のような質問文である。この質問文は「予想最高気温や」と「最低気温は」の文節が並列構造になっているが、最高気温と最低気温の予想方法という１つのことだけを尋ねている。よって、この質問文は単質問文になる。そして、このような並列構造を含む質問文は次の判定処理３によって分類される。

2.3.　判定処理３

処理３では、SVM を利用して複数質問文と単質問文の分類を行うが、その SVM で使用する素性はそれぞれ、又は両方に現れる特徴を主に利用する。質問文に現れる特徴は以下の特徴１から特徴３である。

- **特徴１－並列構造**

[1] http://research.nii.ac.jp/ntcir/workshop/index-ja.html

[2] http://svmlight.joachims.org/

質問文に対して構文解析器KNP[3]を用いて構文解析を行った結果、並列構造を含む質問文が多く見られた。並列構造に関して、菅沼ら[2][3]の研究で並列構造を含む質問文を階層的に分類しているが、本研究では並列構造を名詞が並列している質問文と文が並列している質問文の２つに分類することにする。図２は名詞が並列している質問文を示す。この図から名詞の「メリット」と「デメリット」が並列していることが確認できる。このように、並列構造が名詞からなる質問文のことを「名詞が並列している質問文」と呼ぶことにする。

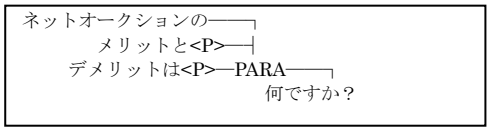

図２　名詞が並列している質問文

図３は文が並列している質問文を示す。この図から「教えて下さい、また」と「合格しましたか？」の両方の文節に関して、動詞を含むことが確認できる。このように、並列構造が文からなる質問文を「文が並列している質問文」と呼ぶことにする。

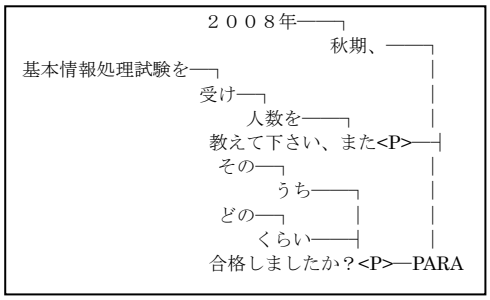

図３　文が並列している質問文

- **特徴２－並立を意味する形態素**

文が並列している複数質問文と単質問文に関して、複数質問文には「並立」を意味する形態素が存在する傾向が有り、単質問文には「並立」を意味する形態素が存在する傾向は無い。並立を意味する形態素とは、本研究では「また」、「又」、「あるいは」、「もしくは」、「ならびに」、「そして」、「かつ」、「さらに」の７種類の形態素を利用する。これら並立を意味する７種類の形態素は、菅沼ら[3]の研究で文が並列している文（完全述語並列）の並列構造を推定する際に使用されている。

- **特徴３－文末距離**

複数質問文と単質問文に関して、並列構造の最後の文節から文末の文節までの距離を「文末距離」と呼ぶことにする。図４はその文末距離の例を示す。そして、文末距離には次の２つの傾向が表れた。

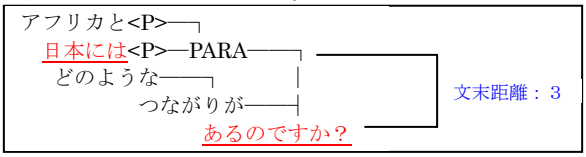

図４　文末距離の例

傾向１　名詞が並列している質問文に関して
「単質問の文末距離」＞「複数質問文の文末距離」
傾向２　文が並列している質問文に関して
複数質問文と単質問文の文末距離は「０」

以上で述べた特徴を考慮し、SVM を利用して分類を行う。そのときに利用する素性は以下の５つである。判定処理３の結果により、複数と判定された質問文の質問数は並列している文節の総数となる。

①	名詞（名詞句・名詞節）が並列しているか
②	文が並列しているか
③	疑問詞表現が存在するか（疑問詞表現）
④	並列構造の文節の形態素に、並立を意味する形態素が存在するか（並立）
⑤	並列構造の最後の文節から文末の文節までの距離（文末距離）

3.　質問数の推定実験

本章では２つの実験とその実験結果について述べる。１つ目の実験は判定処理全体について行い、２つ目の実験は判定処理３で使用する SVM がどの程度判定できるかを調べる。実験１と実験２の実験データとして、単質問文は NTCIR から、複数質問文はインターネットの質問サイト等から任意に取得したものを使用する。

3.1.　判定処理全体の実験（実験１）

この実験では、判定処理１から判定処理３までの精度を調べるために Closed テストと Open テストを実施する。Closed テストでは質問文の一部（判定処理３のみ）に SVM の学習データを使用し、Open テストでは質問文に SVM の学習データを一切使用していない。実験方法に関して、用意した質問文の全てに対して判定処理を行い、判定処理１から判定処理３で分類された質問文を集計し、正しく判定されたかを調べて精度を取る。この実験での精度は以下の式から導く。

$$\text{精度} = \frac{\text{判定処理の正解数}}{\text{質問文の入力数}}$$

表１　実験１の Closed テストの結果

Closed テスト	処理1	処理2	処理3	合計
入力文数	5 文	10 文	85 文	100 文
精度	100.00%	100.00%	90.59%	96.86%

表２　実験１の Open テストの結果

Open テスト	処理1	処理2	処理3	合計
入力文数	3 文	5 文	42 文	50 文
精度	100.00%	100.00%	83.33%	94.44%

表１と表２は実験１の Closed テストと Open テストの結果を示す。表１と表２の入力文数は質問文の数、精度は小数第３位で四捨五入されているパーセンテージを示す。

判定処理１はパターンマッチングを使用して、質問文の形態素列に疑問詞表現が複数含まれていたときに、その質問文を複数質問文とする処理である。実験１の結果より、パターンマッチングのみでも十分に複数質問文の判定を行うことが可能であるということが言える。判定処理２は判定処理１から渡された質問文を構文解析し、並列構造がない質問文を単質問文とする処理である。判定処理１と同様、判定処理２の精度は 100%であったためにこの判定方法も単質問文の判定を行うことが可能であると言える。判定処理３に関しては次の実験で詳しく述べる。

[3] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html

3.2. 判定処理３の実験（実験２）

この実験では、Closed データと Open データを使用して SVM の精度を調べる。実験方法は単質問文のみのテストデータ、複数質問文のみのテストデータを用意し、判定処理を行う。その結果から、正しく判定できているかを調べる。その際の精度の求め方は実験１と同様である。

表３　実験２の Closed テストの結果

使用する SVM の素性	単質問文	複数質問文	全体（平均）
全ての素性	94.00%	92.00%	93.00%
「文末距離」の素性以外	73.00%	62.00%	67.50%
「並立」の素性以外	69.00%	95.00%	82.00%
「疑問詞表現」の素性以外	94.00%	92.00%	93.00%
「文」の素性以外	74.00%	94.00%	84.00%
「名詞」の素性以外	74.00%	94.00%	84.00%

表４　実験２の Open テストの結果

使用する SVM の素性	単質問文	複数質問文	全体（平均）
全ての素性	86.00%	86.00%	86.00%
「文末距離」の素性以外	84.00%	78.00%	81.00%
「並立」の素性以外	64.00%	92.00%	78.00%
「疑問詞表現」の素性以外	86.00%	86.00%	86.00%
「文」の素性以外	66.00%	66.00%	66.00%
「名詞」の素性以外	66.00%	88.00%	77.00%

表３と表４は実験２の Closed テスト、Open テストの結果を示す。全ての素性を使用した場合の精度は、Closed テストに関しては単質問文 94%、複数質問文 92%、全体 93%で、Open テストに関しては単質問文 86%、複数質問文 86%、全体 86%という結果になった。それでは、この結果と素性を１つずつ取り除いた結果を比較する。

まず Closed テストについて述べる。「文末距離」を使用しなかった場合、単質問文が 21%、複数質問文が 30%、全体が 25.5%下がった。任意に取得した複数質問文 253 文、単質問文 120 文の文末距離の平均に関して調べたところ、前者は 1.75 で、後者は 4.09 であった。このことからも文末距離は有効であるといえる。誤判定になった質問文は、例えば、単質問文「WTO と FTA との関係はどうなっているのですか？」のような質問文や複数質問文「映画「タイタニック」の主役とヒロインは誰でしたっけ？」のような質問文があった。

「並立」を使用しなかった場合、複数質問文が 3%上がったが、単質問文が 25%、全体が 11%下がった。単質問文の結果に関して、文が並列している質問文は 20 文存在したが、全て誤判定されていた。複数質問文の結果に関して、３文が誤判定から正しい判定に修正されていた。その中に２文に関しては文が並列している質問文かつ並立を意味する形態素を含んでいない。つまり、複数質問文に関して、「並立」の素性を使用した場合は並立を意味する形態素を含んでいない文が並列している質問文が誤判定になり、「並立」の素性を使用しない場合は上記の質問文は正しい判定になると推測できる。よって、「並立」の素性を取り除いた結果、この２文が正しい判定に修正されたと思われる。

「疑問詞表現」を使用しなかった場合、単質問文、複数質問文、全体、全ての精度は変化がなかった。さらに、誤判定になった質問文も全て同じ結果になった。疑問詞表現の必要性に関して、質問文の最後の表層表現を考えたとき、複数質問文は疑問詞表現を利用せず、単質問文では疑問詞表現を利用する傾向があるのではないかと推測した。例えば、複数質問文では並列している名詞が時間を意味する「タイム」と、人を意味する「記録保持者」の場合、質問文の語尾が「男子１００ｍ走の世界記録のタイムとその記録保持者はだれか教えてください。」のような疑問詞表現を利用しない。それに対して、「ある歌手の名前と年齢が同じ人は誰かいますか？」という質問文を考えたとき、「名前」と「年齢」が並列しているが、「人」を尋ねているためにこの質問文の語尾は疑問詞表現を利用している。すなわち、「疑問詞表現」は役立つと予想していた。しかし実験結果から、この素性は全く機能していない。Closed データを調べてみると、「疑問詞表現」を利用している質問文の平均は複数質問文 0.5 であり、単質問文 0.77 であった。よって、データ次第でこの平均の差が広がれば、「疑問詞表現」も利用できるのではないかと考える。

Open テストの実験結果にも Closed テストと同じような傾向が表れた。

4. 質問タイプの同定の提案

図５は本研究で提案するタイプ同定までの流れ図を示す。判定処理１で複数質問文と判定された質問文に対して、疑問詞表現がその質問文の質問タイプとなる。例えば複数質問文中に「誰」という疑問詞表現があれば、

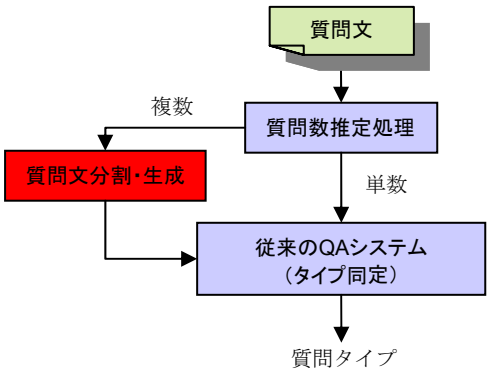

図５　タイプ同定までの流れ図

その質問タイプは人名になるし、「どこ」なら場所で、「いつ」なら時間が質問タイプになる。判定処理３で単質問文と判定された質問文に対して、従来のシステムにそのまま渡し、複数質問文と判定された場合、並列構造にある並列している文節の数に従い、疑問文を分割して従来の QA システムの入力とする。ただし、ここで提案するタイプ同定はまだ実験を行っていないため、実際にどの程度タイプ同定を行うことが可能かは未知である。

図６は質問文分割・生成のアルゴリズムを示し、図７と図８はそれぞれ名詞、文が並列する質問文の分割例を示す。図６中の共通部分とは、「並列構造、又は並列構造より文節番号の大きい文節に係っている一番文節番号が大きい文節とそれよりも文節番号が小さい全ての文節」と「並列構造の最後の文節の次の文節とそれよりも文節番号が大きい全ての文節」のことを指す。図７での共通部分は「ネットオークションの」と「何ですか？」で、図８には共通部分は存在しない。

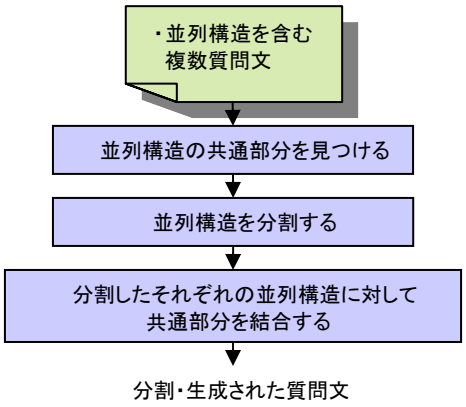

図６　質問文分割のアルゴリズム

それでは、図７中の質問文「ネットオークションのメリットとデメリットは何ですか？」を用いて質問文分割・生成の方法を説明していく。

まず始めに、並列構造の共通部分を探す。図７中の１行目の文節「ネットオークションの」は並列構造に係っていて、この文節よりも文節番号が小さい文節番号が存在しないため、この１行目の文節が共通部分になる。さらに最後の文節「何ですか？」が並列構造の係り先になっていて、この文節よりも文節番号が大きい文節は存在しないため、この最後の文節がもう一つの共通部分になる。よって、１

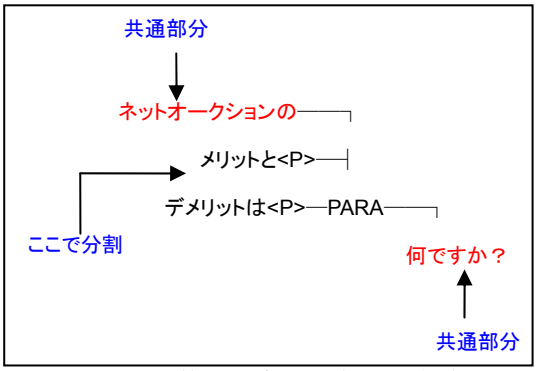

図７　質問文分割の例（名詞が並列）

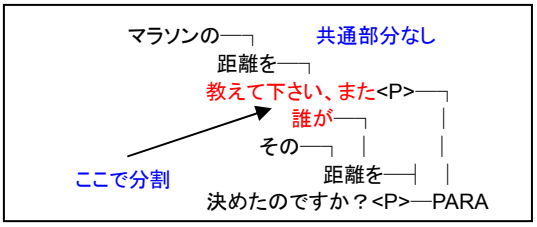

図８　質問文分割の例（文が並列）

行目の「ネットオークションの」と最後の「何ですか？」の文節、この２つが共通部分になる。次に、並列構造を分割する。この質問文中の並列構造は２行目の「メリットと」の文節と３行目の「デメリットは」の文節か成り立っている。よって、この２行目と３行目の文節を分割する。最後に、分割したそれぞれの並列構造に対して共通部分を結合する。すなわち、分割した２行目の文節に１行目と最後の文節を、分割した３行目の文節に１行目と最後の文節を繋ぐ。この結果、「ネットオークションのメリットと何ですか？」と「ネットオークションのデメリットは何ですか？」という単質問文が２文生成される。ここで、前者の単質問文は明らかに日本語として間違った文章であるが、文分割・生成で生成された単質問文はタイプ同定を行うのに使用されるため、たとえ日本語として間違っていてもタイプ同定が正しく判定できればよい。よって、生成された単質問文「ネットオークションのメリットと何ですか？」はタイプ同定が正しく判定できる文章と思われる。

表５　質問文分割・生成の成功例と失敗例の一部

質問文1：ネットオークションのメリットとデメリットは何ですか？
生成文1：ネットオークションのメリットと何ですか？
生成文2：ネットオークションのデメリットは何ですか？
質問文2：放射線のない空間はどのような場所がありますか、また、放射線のない状態は人体に影響を与えますか。
生成文1：放射線のない空間はどのような場所がありますか、また、
生成文2：放射線のない状態は人体に影響を与えますか。
質問文3：鼻血が出る理由とその止血法は何ですか？
生成文1：鼻血が出る理由と何ですか？
生成文2：その止血法は何ですか？
質問文4：世界の中で原子力発電所を持っている国、持っていない国を知りたいです。
生成文1：世界の中で原子力発電所を持っている国、知りたいです。
生成文2：持っていない国を知りたいです。

表５は質問文分割・生成の成功例と失敗例の一部を示す。「ネットオークションのメリットとデメリットは何ですか？」や「放射線のない空間はどのような場所がありますか、また、放射線のない状態は人体に影響を与えますか。」は質問文分割・生成の成功例である。それに対して、「鼻血が出る理由とその止血法は何ですか？」や「世界の中で原子力発電所を持っている国、持っていない国を知りたいです。」等の質問文に関して、前者では「鼻血が出る理由と何ですか？」と「その止血法は何ですか？」の２つに質問文分割・生成される。「その止血法は何ですか？」の質問文は文章中に代名詞の「その」が使用されていることで、「その」にあたる名詞がわからないためにタイプ同定が正しく判定されない可能性が高い。このような質問文に対してタイプ同定を行う場合には代名詞の先行詞を求める照応解析が必要である。後者では、質問文分割・生成を行うと「世界の中で原子力発電所を持っている国、知りたいです。」と「持っていない国を知りたいです。」の２つの文が生成される。「持っていない国を知りたいです。」の文は主語が抜けているためにタイプ同定が正しく判定されないと推測できる。

5.　おわりに

本研究では、質問文中に含まれる質問数を推定し、その結果により質問タイプを同定する手法を提案した。質問数の推定実験では、本手法の有効性を示した結果が得られたが、現時点では、正確な日本語での文の入力のみを考えているので、省略文やくだけた日本語の文などにも対応させる必要がある。また、タイプ同定に関して、今回は手法の提案だけに留まったので、提案した同定手法を使い、実際に同定実験を行う予定である。

参考文献

[1] 田村 晃裕，高村 大也，奥村 学：複数文質問のタイプ同定，情報処理学会論文誌，Vol.47，No.6，pp.1954-1962.（2006）
[2] 菅沼 明，山村 広臣，牛島 和夫：日本語文における名詞句の並列構造の推定およびその推敲支援への適応，情報処理学会論文誌，Vol.38，No.7，pp.1296-1307.（1997）
[3] 菅沼 明，宮崎 晋：日本語文章の推敲支援を目的とした並列構造の指摘，情報処理学会研究報告，DD-99-25，pp.41-48.（1999）