質問応答システムにおけるマニュアルからの技術情報抽出

5V－7

斎藤由香梨、落谷亮、松井くにお、杉山健司
株式会社 富士通研究所

1　はじめに

ＱＡ事例を利用してユーザからの質問に答える質問応答システム [1] では、ＱＡ事例の充実が必要である。事例の充実には、実際のＱＡ事例を収集し、登録することが考えられる。しかし、実際の例を収集するだけでは事例を網羅するには不十分であったり、また、新製品などでは実際のＱＡ事例の蓄積がなかったりする。そこで、我々は、質問応答システムのＱＡ事例を充実させるため、既存のマニュアル等の文書からＱＡ事例となりうる情報を抽出する方式についての研究開発を行なっている。

2　マニュアル文の特徴

マニュアル等の文書から、ＱＡ事例となりうる重要な情報を抽出する方式を検討するにあたり、そのような情報がどのような文で表現されているかを知るため、マニュアルの文章の特徴を調査した。調査したのは、エキスパートシステム構築ツール KwESHELL-KW [2] のマニュアルである。その結果、マニュアル文では、類似した表現を用いていることが多いことが分かった。

例えば、KwESHELL-KW のマニュアルでは、用語の定義を述べる場合、「～を～という」という表現が多く使われている。以下に例文を示す。

- 知識ベース部品を総称して「ナレッジウェア」という。
- クラスが規定するインスタンスの属性をスロットという。
- クラス es:bb のサブクラスを総称して黒板クラスという。

また、KwESHELL-KW で使われるデータの操作方法については、「～は～（によって／により）～される」という表現が多く使用されている。以下に例文を示す。

- 任意のクラスのインスタンスは、関数 make-instance によって生成される。
- インスタンスは，総称関数 es:destroy-object によって消去される。
- メソッドは定義マクロ defmethod によってメソッドクラスのインスタンスとして定義される。

KwESHELL-KW のマニュアルを調査した結果、このような表現が 44 個あった。

このような特徴的な文はＱＡ事例となりうる情報の抽出に都合が良い。

前に示した「～を～という」という文は、「ＸをＹという」という表現にまとめることができ、Ｙの部分が用語名、Ｘの部分がその説明に相当する。この表現を持つ文からは、「Ｙとは何か？」「Ｘのことである」という質問応答例の作成が可能となる。

同様に、「～は～（によって｜により）～される」という文は「ＸはＹ（によって／により）Ｚされる」という表現になり、Ｘが操作される対象、Ｙがその手段、Ｚが操作の内容に相当する。このような表現の文からは、「ＸをＺするには？」「Ｙを用いればよい」というような質問応答例が作成できる。

そこで、我々はマニュアル中の特徴的な表現を調査し、その表現をパターンとして記述し、そのパターンと照合する文を抽出し、ＱＡ事例を作成することにした。

3　抽出方式の比較

前節で述べたようにマニュアルから特徴のある表現を探す場合、単純な方法としては文字列によってパターンを記述して、その文字列を含む文を抽出する方法がある。しかし、単純に文字列で検索したのでは精度が悪い。

例えば「～とは」をという文字列を含む文を検索すると「～することはできる」のような文とも照合してしまう。また、「～と～できる」という表現を含む文を抽出したい場合、「と」の直前の語は動詞の終止形であると記述しないと「～するときに～できる」のような文にも照合してしまう。

そこで、原文を形態素解析し、その結果と形態素の表記や形態素の品詞情報を記述したパターンとを照合することにより、文の抽出を行なった。そして、この形態素を用いる方式と文字列で抽出する方式との比較を行なった。

Automatic Extraction of Technical Information from Manuals
Saitoh Yukari, Ochitani Ryo, Matsui Kunio, Sugiyama Kenji
Software Laboratory, Fujitsu Laboratories Ltd.

表 1: 文字列方式による抽出実験

表現	KW	QAC	CBM	小計
1	1/1	1/1	5/5	7/7
2	6/58	1/16	2/18	9/92
3	17/37	0/1	4/7	21/45
4	22/116	9/15	22/43	53/174
5	8/24	2/4	4/9	14/37
合計	54/236	13/37	37/82	104/355
正解率 (%)	22.9	35.1	45.1	29.3

(正解文数／抽出文数)

表 2: 形態素方式による抽出実験

表現	KW	QAC	CBM	小計
1	1/1	1/1	5/5	7/7
2	6/6	1/2	2/2	9/10
3	17/24	0/0	4/5	21/29
4	22/38	9/9	22/23	53/70
5	8/14	2/2	4/4	14/20
合計	54/83	13/14	37/39	104/136
正解率 (%)	65.1	92.9	94.9	76.5

(正解文数／抽出文数)

比較実験の対象のマニュアルとして富士通（株）発行の三つのマニュアル、KwESHELL-KW(KW), 質問応答システムであるＱＡクライアント (QAC)、事例管理システムＣＢマネジャー (CBM) のマニュアルを用いた。ここで使用した特徴的な表現は、以下の５種類である。

1. ～とは～である
2. ～ |動詞終止形| と～できる
3. ～は～（で｜でも｜によって｜による｜を用いて）～できる
4. ～ |動詞終止形| （には｜ために｜ためには）～
5. ～は～ |動詞終止形| ために（の｜に）～

表１は、文字列を用いて文を抽出した場合の文数とその内の正解数である。表２は、形態素の表記や品詞情報を記述したパターンを用いて抽出した結果である。

例えば、KwESHELL-KW のマニュアルから「～と～できる」という文字列を用いて抽出した場合、58 文が抽出され、その 58 文中、正しいものは６文、誤りが 52 文であった（表１の表現２の KW の項参照）。これに対して、形態素の表記とその品詞情報を記述したパターン「動詞終止形／と／～／でき／る」を用いると６文が抽出され（表２の表現２の KW の項参照）、その６文は全て正しかった。このように形態素の表記や品詞情報を用いる方式は、文字列を用いる方式と再現率は同じで、全体の正解率は 29.3% から 76.5% にアップしている。

4 抽出実験

KwESHELL-KW, ＱＡクライアント、ＣＢマネジャーのマニュアルに対して、形態素の表記や品詞情報を記述したパターンを 105 個を作成し、文の抽出実験を行なった。各マニュアルの文の数は、KwESHELL-KW は 2165 文、ＱＡクライアントは 255 文、ＣＢマネジャーは 252 文である。

表 3: 抽出結果

マニュアル名	抽出文数	正解数	正解率 (%)
KW	874	796	91.1
QAC	140	130	92.9
CBM	144	135	93.8

文を抽出した結果が表３である。ここで、正解数は抽出文数中の正しく抽出された文の数、正解率はその割合を示している。抽出した文の正解率は 90 %以上であり、この方式は文の抽出に有効であると考えられる。

5 今後の課題

形態素の表記や品詞情報を記述したパターンを用いても、誤った文が抽出されるのは、主に意味的に誤っている文を抽出してしまう場合である。今後は品詞情報の他に意味の情報も利用できるようにする必要がある。

また、今回の抽出実験では、抽出する単位を一文としている。しかし、一文で内容が完結していない場合も多い。その場合には、一文だけではなく、関連する箇所を同時に抽出する必要がある。今後はこのように情報のまとまりをブロックとして抽出することに取り組んでいく予定である。

さらに、こうしてマニュアルから抽出した情報が実際の質問応答システムに組み込まれた場合に、ユーザにとってどの程度有用であるかについての評価も行う予定である。

参考文献

[1] 佐藤他: 事例データベースを利用した質問応答システムの構築, 情報処理学会第 49 回全国大会 5V-06(1994)

[2] 富士通株式会社, エキスパートシステム構築ナレッジウェア KwESHELL-KW(1991,1992)