E-013

表構造を利用した質問応答システム

Question Answering System Using Table Structure

大久保　和之†　　鈴木　康広†
Kazuyuki Okubo　Yasuhiro Suzuki

１．まえがき

近年、自然言語による質問応答システムの研究が盛んに行われている。我々は、質問文とその回答を含むと思われる文章中の助詞や述語などに着目し、文章中の主語、目的、理由、場所などを表にまとめ、それらの情報を基に回答を得る質問応答システムの研究を行っている(1)。本報告では、実験システムの概要と評価について述べる。

２．表構造を利用した質問応答

我々は質問応答の際に表構造を用いることで回答を得る実験を行った。表構造とは、以下に示すように文章の助詞や述語などの表層構造に着目し、表として扱う手法であり、表構造を使用する事で、深い文法的な解析や文法情報を使用することなく質問応答を行う。

例）私は大学で勉強した

主語	私
場所	大学
行為	勉強した

質問文とその回答を含むと思われる文章の形態素解析結果から上記の表（表構造）を作成し、それらを比較することで回答を得る質問応答手法を提案する。

３．実験システム

提案した手法の有効性を確認するために実験システムを構築し、評価実験を行った。

実験システムの処理の流れを以下に示す。以下、図１に従って実験システムの概要について述べる。

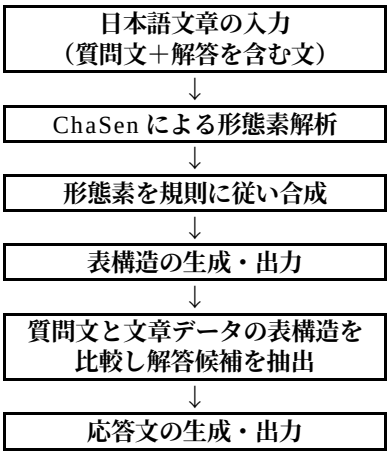

図１　実験システムの処理の流れ

†北海道工業大学大学院　電気工学専攻 1

3.1 形態素の合成

入力された質問文と文章データは、まず形態素解析ツール（ChaSen[1]）を用いて形態素解析を行い形態素に分割される。その後、形態素同士の品詞と合成規則との比較を行い、形態素合成を行う事で、解答候補として出力するために意味のある単語にする。

例えば、文章中に「ハリウッド･スター」という単語が出現した場合、形態素解析を行うと「ハリウッド」＋「･」＋「スター」と分割されてしまうが、規則にのっとって合成を行う事で「ハリウッド･スター」と一つの単語にまとめることが出来る為、解答候補として出力出来る形となる。

3.2 表構造の生成・出力

本システムで用いる表構造とは、文章中の単語を、主語・述語・目的語に分類した表の事で、各単語の役割を表面化させ管理する事で、解答候補の抽出を行う際に利用する情報となる。

主語･述語･目的語の分類は文章中の述語に着目して行う。例えば、私は～　という形で係助詞「は」が現れた場合、助詞の前の単語「私」は主語である事が分かる。こういった、助詞と単語の分類を纏めた規則を生成し、作成した規則と照らし合わせることで、表構造の生成を行う。

入力された質問文と解答を含む文それぞれの表構造を作成し、解答候補抽出の際に利用する。

3.3 解答候補の抽出

解答候補を抽出する際には、初めに、質問文の表構造中の質問を表す単語を抽出し、質問文の分類を行う。
質問文の分類は 5w1h の 6 パターン（What When Where Why Who How）に How much を加えた 5w2h の７種類を用いる。

分類した質問文のジャンル情報に加え、解答を含む文中の各単語毎の主語･述語･目的語の分類情報と形態素解析の際に得た単語の品詞情報の３つを使用し、それぞれの単語に解答候補としてのスコアを付ける。

最終的に、単語につけられたスコアの高い物を解答候補として抽出し、出力する。

４．性能評価実験

表構造を利用した質問応答の性能評価を行うため、質問文１５３文と、各質問に対して、解答を含む文章をインターネットの web サイト上から１～１０文（合計 994 文）用意し、表 1 の条件の下で実験を行った。

実験はまず、手動での表構造生成を全質問文とその解答を含む文に対して行う事で、形態素の合成と表構造の生成に試用する規則を生成した。その後、生成した規則を利用し質問応答の実験を行った

表１　実験に用いた条件

質問文データ数	１５３文
テキストデータ	９９４文 (各質問文に対し1～１０文)
使用形態素解析ツール	ChaSen[1]
使用検索エンジン	google[2]

4.1 文章情報処理規則の生成

解答を含む文章は

検索サイト google[2] を用いて、解答となる単語を含む文章を検索するが、その際に用いる検索用キーワードの決定には文章の解析を行う必要がある。しかし今回の実験の目的は、表構造を用いたシステムの評価を行うことにあるので、質問文を解析しキーワードを決めるのではなくアンケートによる集計結果を利用し、決定した。

アンケートでは１５３の質問文の解答を求めるためには、質問文中のどの単語を検索用キーワードとして利用するかを質問文に印を付ける形で回答させ、一つの質問文に複数の検索単語があっても良いとした。

アンケートの集計は質問文の各形態素毎に検索単語として用いられているかを集計することで、質問文中のどの形態素が多く解答の検索用語として用いられるかを集計した。アンケートは合計 12 人の被験者に対して行い集計を行った。

集計結果から得られたキーワードのうち、各質問文中の検索用に用いられる率の高い上位３つのキーワードでの検索を行った結果、上位 3 位に出現したサイトから解答を含むテキストとその前後のテキストを抜き出し、テキストデータとした。

4.２ 文章情報処理規則の生成

(1) 形態素の合成に使用する規則を生成するため､手動で単語合成を行い、合成する必要のある述語の組み合わせのパターンをリストとして生成した。

その後、作成したリストを規則として、単語の合成を行い、単語の合成が上手く行われない規則の候補の削除・新たに必要になった規則の追加を繰り返し､最終的に 30 の単語合成規則を生成した。単語合成規則の例を表 2 に示す。

表２のうちｘ列はある単語の品詞を示し、ｘ＋１はその次の単語の品詞を示す。

二つの単語の品詞が規則と一致した場合、その二つの単語は合成され、品詞情報は「合成後」の列に書かれた品詞となる。

表２　単語合成規則（一部）

ｘ	ｘ＋1	合成後
名詞	名詞	名詞
動詞	名詞	名詞
動詞	動詞	動詞
助詞	助詞	助詞
接頭詞	名詞	名詞

[1] http://chasen.naist.jp/hikki/ChaSen

[2] http://www.google.co.jp

(2)表構造の生成を行うための規則を抽出するため、手動で表構造の作成を行い、作成した表構造の情報を集計し、各単語の表層語と品詞の組み合わせとし、出現回数が一定以上のものを表構造生成のための規則として試用することとした。

結果として、規則の数は 619 となった。表構造作成規則の例を表 3 に示す。テキストデータから表構造を作成する場合表生成規則とテキストの単語を比較し、一致するものを探して表構造を作成するが、規則のどれも一致するものが無かった場合には、表構造は「不明」と出力される。

表３　表構造生成規則（一部）

表構造	品詞	助詞	助詞種類
目的	名詞	の	助詞-連体化
目的	名詞	を	助詞-格助詞-一般
主語	名詞	は	助詞-係助詞
主語	名詞	の	助詞-連体化
主語	名詞	が	助詞-格助詞-一般
目的		として	助詞-格助詞-連語

4.３ 質問応答実験

質問応答を行う際には、出力する解答の候補は、単語のスコアの最も高い物を出力する事としたが、同スコアのものが幾つか現れる場面が多かったため、同じスコアであった場合、回答は３つまで候補を出力する事とした。

実験によって得られる回答の正解の判定は正解･準正解の 2 段階とし、解答候補が一つで尚且つそれが正解の場合が正解、回答候補が二つ以上であるが、その中に回答が含まれる場合は準正解であるとした。表 4 に実験での正答率を示す。

表４　実験結果　正答率（%）

正解	準正解	不正解
3.4	81.6	15.0

5．まとめと課題

実験により、正解・準正解の合計では８５%と高い精度での結果を得ることが出来た。これは、ある程度まで解答候補の絞込みを行う事は出来ると考えられる。

しかし解答を一つに絞るという意味では、３%程度の正答率になってしまった。

理由として考えられるのが、解答候補の選定に対する条件の不足であり、今後はより多くの条件で解答候補のスコア付けをする事によって、より高い精度での解答抽出を行うことが出来るようにしたい。

参考文献
(1)岡崎 崇宏、米谷 順司
北海道工業大学　電気工学科卒業論文（平成 14 年度）
「表構造を利用した質問応答ステムの研究」