ウェブからの情報抽出システムの構築
－定義型質問に対する情報検索に基づく回答の作成－

川浪理恵子† 大熊智子‡ 増市博‡ 杉原大悟‡ 石崎俊†

†慶應義塾大学政策・メディア研究科 ‡富士ゼロックス(株)研究本部

†慶應義塾大学環境情報学部

1. はじめに

我々は日常、知らない言葉に出会うことがよくある。知らない言葉を調べる手段としては、辞書を引くということが一般的であるが、新しい言葉や専門的な言葉などは一般的な辞書には収録されていないので調べることができない。一方、「ウェブで調べる」という方法では、大量のウェブページを検索することが可能なので、辞書に収録されていないような言葉の意味を調べることができる可能性が高い。しかし、現在利用可能なサーチエンジンの多くは、検索語を入力すると検索語を含むウェブページのＵＲＬが複数示されるようになっており、大量のページの中から手作業で知りたい言葉の意味が書いてある部分を見つける必要があるため、ユーザにとって負担が大きい。

文書検索の次の段階として、質問応答システムがある。これは、質問を入力すると回答が出力されるもので、文書検索システムよりも知りたいことへ辿り着く時間が短い。しかし、現在、名称や数値を尋ねる factoid 型質問に回答するシステムは存在するが、言葉の定義を尋ねる definition 型質問に回答するシステムはあまりない。その理由は、第一に、factoid 型は固有表現抽出ができれば回答できるが definition 型は回答が単語ではなく句や節の形であり回答抽出が難しいこと、第二に、factoid 型は唯一の正解が存在するので評価が容易であるが、definition 型は正解がひとつに定まらないので評価が難しいことである。

従来、テキストから用語の定義を抽出する研究として、パターンマッチングに基づいた手法が提案されている[1][2][4][5][6]。西野らは、「αとはβである」という表現に的を絞り、新聞記事からの用語と定義を抽出している。また木田らは「αはβ」という表現を分析し、αとβの関係を記述している。本研究では、これらの研究を基にして、用語を入力すると、パターンマッチング手法によってウェブ文書からその定義となるフレーズを抽出し、回答として出力する質問応答システムを実現した。また、回答の有用性を示すために、文章を読んでその中の知らない言葉の意味を調べるタスクを設定し、ユーザによる評価実験を実施した。

2. システムの概要

本システムの全体図を図 1に示す。

質問（定義を知りたい語）を入力すると、まず検索式構成部で入力された語に拡張パターンを付加した検索式が作成される。次に、文書収集部で作成した検索式をウェブサーチエンジン（goo[7]）にかけ、ウェブページを収集する。そして、情報抽出部で、抽出パターンを用いて、パターンマッチングにより、ウェブページ中から回答候補を複数抽出する。最後に、回答選択部で、抽出に用いた抽出パターンの優先順位をもとに、最も確からしい回答をひとつ、あるいは複数選択して出力する。

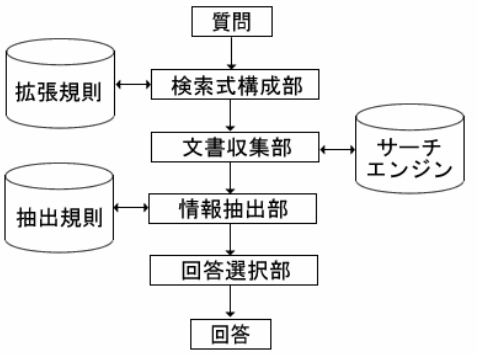

図 1：システム構成

3. 定義型質問に対する望ましい回答とは

システムの詳細を述べる前に、まず、本システムが出力すべきものは何かということを明らかにしておく必要がある。

本システムは質問応答システムであり、出力は回答そのものである必要がある。factoid 型の質問に対しては、何らかの名称あるいは数値を出力すればよいが、定義型の質問に対する望ましい回答がどのようなものであるかという答えを見つけるのは難しい。

長尾[3]のいう「内包的定義」は、「ある物事の集合に共通な、しかもそれによってそれらの物事がほかの物事から明瞭に区別され得るような本質的な諸特徴」を示すものである。これは百科事典の用語説明の第１文に典型的に見られるものであり、本システムの回答としてふさわしいものと言える。しかし、「辞書・事典の読者に用語を十分理解させるためには、できるだけ多くの観点から定義を行うことが望まれる」ため、本システムにおいては、「内包的定義」を中心に、さまざまな観点からの定義を出力することを試みた。

4. 定義文抽出の手法

本システム中で最も重要なのは、パターンマッチングにより、ウェブページ中から回答候補を抽出する部分（情報抽出部）と、それらの回答候補から最も確からしい回答を選択する部分（回答選択部）である。ここでは、情報抽出部、回答選択部における処理の流れを説明する。

まず、区切り文字（「。」、「！」、「？」）を手掛かりに、テキストを文単位に区切る。次に、それらの文の中から、入力された語にマッチする部分（以下αと呼ぶ）を含む文を見つける。そのような文が見つかったら、文を chasen[8]を使用して形態素解析、あるいは cabocha[9]を使用して構文解析し、作成した抽出パターンとのマッチ

ングを行う(形態素解析の結果、αの直前または直後に、接頭詞、名詞、あるいは未知語が存在した場合、または構文解析の結果、αを含むチャンクに係るチャンクが存在した場合は、文字列αを含む複合名詞や、特定のαに関する文である可能性が高いので除外する)。抽出パターンは優先度によって 12 段階に分かれている。抽出パターンの例を表 1に示す。なお、αは入力された文字列、βは任意の文字列を示し、{A/B/C}は、A、B、C のいずれかの要素を示す。

表 1：抽出パターンの例

1. αとはβを{指す/言う/意味する}
2. αとはβという意味{だ/です/である}
3. αとはβ{こと/もの}{だ/です/である}
4. αとはβ{だ/です/である} αはβ[名詞]{。/！/？}
5. αはβを{指す/言う/意味する}
6. αはβという意味{だ/です/である}
7. αはβ{こと/もの}{だ/です/である}
8. αはβ{だ/です/である}
9. αはβ[名詞]{。/！/？}
10. βをαと呼ぶ βはαと呼ばれること{も/が}ある
11. β{「/『/“}α{」/』/”} βがα{だ/です/である}
12. α（β）

これらのパターンでマッチングを行った結果、βにマッチした部分を回答候補とする。ただし、代名詞を含むもの、3 文字以下のもの、数字と記号のみから成るものは除外する。使用したパターンの優先度順に回答候補をソートし（同一優先度内では、文字数の多いものを優先する)、優先度の高いものから順に出力する。

5. 評価実験

5.1. 評価実験の動機

本システムの動作及び有用性を調べるために、(1)回答選択の精度を調べる実験、(2)文中の言葉の意味を調べるタスクを行った。(1)は回答の正しさを、(2)は回答の有用性を調べるためのものである。

実験の詳細を説明する前に、このようなふたつの実験を設定した理由を述べる。

definition 型の質問（「～とは何ですか」など）は正解がひとつに定まらないため、評価が難しい。その理由は、第一に、用語はさまざまな観点から定義することができるため、第二に、ユーザの目的によって、求められるものが異なるためである。本システムは、手早く用語を調べることを目的とするユーザを想定しており、「簡潔で役に立つ、正しい」回答を出力することを目的としている。「簡潔さ」については、回答のボリュームが小さければ簡潔である、と仮定し、これを回答を全て一文で出力することにより達成しているので、評価は、「正しいかどうか」と「役に立つかどうか」の観点から行えばよい。

「正しいかどうか」については、用語がさまざまな観点から定義可能であることを考えると、何をもって「正しい」とするのかが明確でないため、正しさを正確に測るのは容易ではない。しかし、便宜的に、内容が間違っていないものを「正しい」と考えるというようにすることはできる。しかし、「役に立つかどうか」は、何を指標に測ればよいのか明らかではない。被験者にシステムの出力した回答を見せて、「役に立つかどうか」を評価してもらうことは可能かもしれない。しかし、「何の」役に立つのかが明確でないと、「役に立つかどうか」の基準には個人差が出てきてしまうと思われるので、この方法が妥当であるかは疑問である。

そこで、本研究では、「新聞、雑誌、本、ウェブページなどの文章を読んでいるとき、わからない言葉があった。わからないまま読み進むよりは、その言葉を調べてから読み進みたい」という状況を設定し、その状況において、システムの出力する回答が役に立つかどうかを評価してもらうタスクを設定することにした。この方法ならば、被験者に、「文章を読み進めるために知らない用語の説明を調べたい」という動機が生まれるので、システムの出力する回答が役に立つかどうかを、実感から判断できるはずである。

5.2. 回答抽出の精度を調べる実験

実験には、『イミダス』（2005 年度版、2002 年度版)、各種学術用語辞典、高校教科書、『大辞林』、goo 注目ワード[1]（2005 年 2 月分、10 月分）から抽出した 250 語の用語を用いた。辞典から選んだ用語は、日本語教師 1 名に、「各辞典からその辞典特有の特徴的な語を選択する」ように依頼し、抽出したものである。各種学術用語辞典は本文を参照しながら、高校教科書は索引から抽出作業を行った。表 2 に抽出した語の例を示す。

表 2：実験に使用した語の例

地獄蒸し，雑穀ソムリエ，ブログ力士，ロボカフェ地下鉄男，共依存症，コンビニ検診，ハクション議連親子留学，サラリーマン川柳，バーチャル内視鏡金利平価，非営利法人，水冷パソコン，ＩＲ育児・介護休業法，パラサイト・シングル，安価ソフト環境権，不確定性原理，青息吐息，大天目，賂ろくろ，水引，クロックムッシュ，鵺，遊水地，知友寄席，紅指し指，制度学派，カーキー選挙，宗教会議相対売買，所有本能，古文辞派，平均律，ピエタディスクキャッシュ，アケメネス朝，廃仏毀釈ステップ気候区，特例公債，電子商取引，歌枕力積，ペプチド結合，解糖系，無限級数

これら 250 語をシステムに入力して実験を行った。結果を表 3 に示す。

ここでは、出力した全ての回答を、「内包的定義」、「属性」、「言い換え」、「不正解」のいずれかに分類し、「内包的定義」、「属性」、「言い換え」に分類されたものを「正解」としている。「内包的定義」には、百科事典などの用語説明の第 1 文と比較し、類似していると認められるもの、「属性」には、百科事典などの用語説明の第 1 文とは内容が異なるが、語の説明として間違っていないもの、「言い換え」には、語を言い換えたもの、外国語に翻訳したもの（または外国語を日本語に翻訳したもの)、漢字

[1] goo が提供する、話題の言葉を紹介するコンテンツ

の読みを分類した。第４位と第５位は正解が存在しなかったので、表では省略した。

入力した 250 語のうち、183 語（73.2%）に対して回答が出力された。回答を出力した 183 語のうち、第１位に正解が現れたものが 137 語、精度（precision）は 74.3%であった。回答を優先度第５位まで出力したときの MRR（Mean Reciprocal Rank）は 0.58 であった。

表 3：評価結果

入力用語数			250
回答出力用語数			183
回答出力率			73.2%
第１位正解用語数	137	内包的定義	102
		属性	23
		言い換え	12
第２位正解用語数	13	内包的定義	7
		属性	5
		言い換え	1
第３位正解用語数	4	内包的定義	0
		属性	2
		言い換え	2
第１位の精度			74.9%
第２位までの精度			82.0%
第３位までの精度			84.2%
MRR（１位～５位）			0.58

回答出力率：回答出力用語数／入力用語数
精度：正解用語数／回答出力用語数
MRR：**最初に正解が表れた順位の逆数を平均したもの**
(NTCIR**の**MRR**平均は**0.3～0.4**)**

74.9%という精度はそれほど高いものではない。これは、今回使用した抽出パターンに精度の高いものと低いものが混在しているからであると考えられる。抽出パターンごとの精度を調べてみると、優先度 7、8、9、11、12 のパターンの精度が低いことがわかった。これらのパターンを不採用にすると、第１位の精度が 74.9%から 87.3%に上がるほか、第２位、第３位までの精度及び MRR が改善された。図２に、全パターンを採用したときと、これらのパターンを不採用にしたときの精度と MRR の比較を示す。

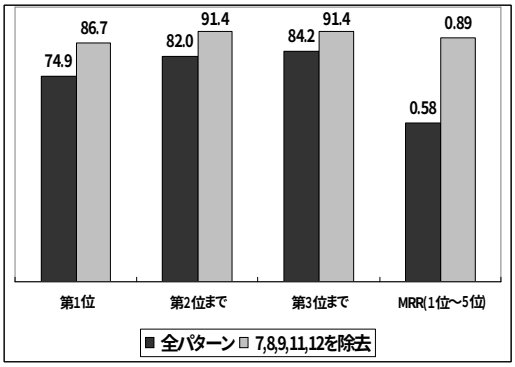

図 2：精度と MRR の比較

NTCIR　QAC-1 では、回答候補を優先順位をつけて５位まで返すというタスクにおいて、全システムの MRR 平均が、0.3～0.4、最高は 0.61 であった。それに対して、本システムの MRR は、0.58 である。難易度の高い definition 型であること、雑音の多いウェブ文書を検索対象文書としていることを考えると、本システムは十分な性能を持っていると言える。

5.3. 文中の言葉の意味を調べるタスク

このタスクは、「新聞、雑誌、本、ウェブページなどの文章を読んでいるとき、わからない言葉があった。わからないまま読み進むよりは、その言葉を調べてから読み進みたい」という状況を設定し、その状況においてシステムの出力する回答が役に立つかどうかを評価してもらうというものである。タスクに使う語は、大辞林、イミダス２００５年度版、goo 注目ワード２００５年１０月分、各種学術用語辞典から選んだ６１語をシステムに入力し、正解を出力した３４語から３０語をランダムに選んだ。被験者が読む文章には、この３０語を検索エンジン（goo）にかけ、ヒットした文書を上から順番に調べていって最初に見つかった「用語を含み、かつ文脈からは用語の意味が推測できないと思われる段落」を使った。被験者は慶應義塾大学総合政策学部４年、環境情報学部４年の学生計 30 名である。表４に実験に使用した回答と文章を示す。

タスクは以下のような手順で実施する。まず、被験者は、ある用語を含む文章を読む。そして、文章全体の意味を理解するのに、その用語の説明があったほうがいいか、なくてもいいかを答える。回答は、「あったほうがいい」あるいは「なくていい」のどちらかを選択する形で行う。ここで、「あったほうがいい」、を選択した被験者には、その用語の説明として、システムの出力を示す。そして、再び文章を読んで、システムの出力が、文章全体の意味を理解するのに役に立ったかどうかを、「十分役に立った」、「ある程度役に立った」、「どちらとも言えない」、「あまり役に立たなかった」、「全然役に立たなかった」の５段階で評価する。「なくていい」を選択した被験者は、すぐに次のタスクに移る。

表 4：実験に使用した回答と文章

用語	回答	文章
IR	アナリスト及び機関投資家等に対して、その会社を正当に評価してもらうための広報活動	企業のインターネット上のIR活動の重要性に関して最近よく耳にします。私たちもお客様からIRサイトのコンテンツに関して相談を受けることがあります。しかし、残念ながら、多くの企業ではまだまだIR活動の意味、IR活動の環境、対象者を本当に理解した上で、自社のIR戦略、IR活動を行なっているとは言えないと思います。
色聴	音を聞いて色を感じる現象のこと	シンポジュウムの二日目。研究発表の中に、色聴の研究をしている人の発表があった。自分自身も色聴の能力があるそうだ。色聴は、女性に多いらしい。

タスクの結果を集計したものを表５に示す。「はい」、「いいえ」は、「用語の説明があったほうがいいですか」

に対する回答、「５」、「４」、「３」、「２」、「１」は、それぞれ「十分役に立った」、「ある程度役に立った」、「どちらとも言えない」、「あまり役に立たなかった」、「全然役に立たなかった」を示す。「平均」は、「十分役に立った」を５点、「ある程度役に立った」を４点、「どちらとも言えない」を３点、「あまり役に立たなかった」を２点、「全然役に立たなかった」を１点として、それぞれの人数を点数と掛け合わせたものの和を、「はい」の人数で割った平均点である。「平均」の列以外は、単位は「人」である。

表 5：タスクの結果

	はい	いいえ	5	4	3	2	1	平均
q1	25	5	10	10	1	4	0	4.04
q2	20	10	14	3	2	1	0	4.50
q3	7	23	1	1	0	5	0	2.71
q4	16	14	0	11	4	1	0	3.63
q5	4	26	0	2	1	0	1	3.00
q6	6	24	4	0	0	1	1	3.83
q7	26	4	12	12	0	2	0	4.31
q8	19	11	9	8	1	1	0	4.32
q9	14	16	7	2	2	3	0	3.93
q10	16	14	13	2	0	0	1	4.63
q11	21	9	6	7	2	4	2	3.52
q12	23	7	6	3	1	8	5	2.87
q13	13	17	5	2	2	3	1	3.54
q14	13	17	7	6	0	0	0	4.54
q15	21	9	7	10	3	0	1	4.05
q16	19	11	11	4	2	2	0	4.26
q17	18	12	12	4	1	1	0	4.50
q18	11	19	1	3	6	1	0	3.36
q19	19	11	7	10	1	1	0	4.21
q20	8	22	2	2	3	0	1	3.50
q21	5	25	2	1	1	1	0	3.80
q22	22	8	5	9	5	3	0	3.73
q23	5	25	1	1	0	2	1	2.80
q24	22	8	5	10	0	4	3	3.45
q25	20	10	5	9	2	3	1	3.70
q26	18	12	9	6	1	2	0	4.22
q27	12	18	5	6	0	1	0	4.25
q28	22	8	13	9	0	0	0	4.59
q29	11	19	6	3	2	0	0	4.36
q30	23	7	2	3	3	10	5	2.43
平均	16.0	14.0	6.2	5.3	1.5	2.1	0.8	3.88

全体の平均は３．８８点であり、システムの出力が、「新聞、雑誌、本、ウェブページなどの文章を読んでいるとき、わからない言葉があった。わからないまま読み進むよりは、その言葉を調べてから読み進みたい」という状況で、「ある程度役に立つ」ことが示された。

平均点４点以上のタスクが存在するのに対し、平均点が２点台のタスクがいくつか存在する。得点の低いタスクの特徴は、①回答が短すぎて含まれる内容が少ないので、文章を理解するために必要な知識としては不足である、②回答が用語のある一側面のみの説明であるため、文章を理解するために必要な情報と回答に含まれる情報が一致しない、③文章中にキーワード以外にも被験者のわからない言葉が存在する可能性がある、④文章の論理構造が難解である、の４点である。今回実施したタスクでは、被験者には第１位に出力された回答のみを示したため、全体として情報量が少なかったことが考えられる。また、タスクを設定する際に、被験者が読むキーワードを含む文章の質の統一を行わなかったため、キーワード以外にもわからない言葉があるなどの理由で文章全体の理解が困難になったタスクが存在したとも考えられる。今後は、第５位までに出力された回答を全て示すなど回答の提示の仕方を含めてシステムを改良する、またタスクに用いる文章としてどのようなものがふさわしいのかという基準を定め、文章のスタイルや内容が文章の理解に与える影響を考慮するなどの必要がある。

6. おわりに

本研究では、「αとはβ」、「αはβ」、「βをαと呼ぶ」などのいくつかの文型パターンを利用して、ウェブから言葉の定義を述べている部分を抽出することにより、回答を作成し、言葉の意味を尋ねる definition 型質問に回答するシステムを構築した。その結果、回答をひとつのみ返す場合の精度はそれほど高いものではなかったが、回答を５つ返す場合には、その中に正解が含まれる割合が大きかった。また、作成したパターンには、強力なものとそうでないものがあることが明らかになった。さらに、タスクにより、本システムの有用性を調べ、「(新聞、雑誌、本、ウェブページなどの)文章を読んでいるとき、わからない言葉があった。わからないまま読み進むよりは、その言葉を調べてから読み進みたい」という場合に、本システムがある程度役に立つことがわかった。

今後の課題としては、強力でないとわかったパターンを改良し、構文情報をより活用したパターンにしていくこと、パターンと抽出できた回答の特徴の相関を調べ、ユーザがどのような観点からの回答を求めているのかによって、特徴の異なる回答を出力できるようにすることなどが挙げられる。

文　献

[1] 木田敦子，乾裕子，落合亮，西野文人：新聞記事からの用語集作成のためのテキスト分析，情報処理学会研究報告，NL134-12，pp.85-92，1999

[2] 桜井裕，佐藤理史：ワールドワイドウェブを利用した用語説明の自動生成，情報処理学会論文誌，Vol.43，No.5，pp.1470-1480，2002

[3] 長尾真：辞典形式での専門分野の知識の体系的構成法，人工知能学会誌，Vol.7，No.2，pp.320-328，1992

[4] 西野文人，橋本三奈子，落合亮：テキストからの用語とその定義文の抽出，言語処理学会第５回年次大会，pp124-127，1999

[5] 藤井敦，石川徹也：World Wide Web を用いた事典知識情報の抽出と組織化，電子情報通信学会論文誌 D-Ⅱ，No.2，pp.300-307，2002

[6] 山田一郎，住吉英樹，柴田正啓：ニュース記事に出現する用語と説明文の意味関係自動獲得，情報処理学会研究報告，NL152-21，pp145-152，2002

[7] http://www.goo.ne.jp

[8] 松本裕治他:”日本語形態素解析システム『茶筌』version 2.2.1 使用説明書", 2000

[9] 工藤拓，松本裕治:"チャンキングの段階適用による日本語係り受け解析," 情報処理学会論文誌，Vol.43，No.6，pp.1834-1842，2002.