Vol.2011-ICS-163 No.2
2011/3/23

質問文の種類に応じた web 検索による単語回答システム

山村伊織[††]　吉村枝里子[†]　土屋誠司[†]　渡部広一[††]

本稿は，質問文の種類に応じた web 検索による単語回答システムに付いて述べている．このシステムは入力された質問文の答えを web から獲得する．また，このシステムは，質問文の求める回答の種類を質問文中から抜き出した場合と，疑問詞から決定した場合で処理を変更している．質問文の求める回答の種類に応じて処理を変更することで，あらゆる質問文に対して質問文の求める回答の種類を正しく特定できるシステムを実現する．

Question and Answer System Using Kinds of Question Sentence Based on The Web Retrieval Method.

Iori Yamamura[††]　and Eriko Yoshimura[†]
and Seiji Tsuchiya[†]　and Hirokazu Watabe[††]

In this paper, it proposes a question and answer system using kinds of question sentence based on the web retrieval method. This system gets answer that was inputted question sentence from web. This system changes processing depending on kinds of the word that was required in the question sentence. This paper proposed the question and answer system can identify correct kinds of the word that was required in the question sentence.

1. はじめに

近年，インターネットの急速な普及，ユーザ数の劇的な増加などの理由から，Web には膨大な情報が存在するようになった．さまざまな情報が電子化された結果，Web 上には新聞記事のようなひとつのテーマについて書かれた文書のみではなく，blog のようにはっきりとしたテーマが無く，信頼性の低い文章も存在している．そのため，既存の検索システムでは膨大な量の情報によって必要な情報が埋もれてしまい，必要な情報を得ることが困難な状況になった．そこで，ユーザが必要としている情報のみを的確に獲得するための技術が求められている．

本研究では，既存システムである知的 web 検索方式[1]を改良して，質問文に対する単語回答を Web から獲得する手法を提案する．このシステムは質問文の種類に応じて処理を変更することで，既存システムの問題点であった質問文の求める回答の種類を正確に決定するものである．本研究では，連想メカニズムを構成する概念ベース[2]と関連度計算方式[3]，質問文意味理解システム[4]，検索エンジン[5]，Web 情報を利用する技術[6] [7]を用いる．本研究では，固有名詞が解答となる質問文に対して柔軟に対応できる手法を実現する．

2. 知的 Web 検索方式

知的 Web 検索方式は，質問文に対する単語回答を Web から獲得する手法である．本研究では，知的 Web 検索方式に 2 点の改良を加えて単語回答システムを作成した．1 点目は質問文の種類による処理の変更であり，この手法で質問文の求める回答の種類をより正確に特定できるようになった．2 点目はスコア付けに検索語と回答候補の Web の共起頻度を使用することであり，これを利用することでシステム全体の精度向上を実現した．

3. 連想メカニズムと質問文意味理解システム

3.1 概念ベース

概念ベースとは複数の国語辞書や新聞などから機械的に構築した，語（概念）とそ

† 同志社大学理工学部
Department of science and engineering, Doshisha University
†† 同志社大学大学院工学研究科
Graduate School of Engineering Doshisha University

の意味特徴を表す単語（属性）の集合からなる知識ベースである．概念と属性のセットにはその重要性を表す重みが付与されている．概念ベースには，現在約 12 万語の概念が収録されており，1 つの概念あたり約 37 個の属性が存在する．

ある概念 A は属性 a_i とその重み w_i の対の集合として式(1)で表される．

$$A = \{(a_1, w_1), (a_2, w_2), \cdots, (a_m, w_m)\} \quad (1)$$

任意の 1 次属性 a_i は，その概念ベース中の概念表記の集合に含まれている語で構成されている．したがって 1 次属性は必ずある概念表記に一致するので，さらにその 1 次属性を抽出することができる. これを概念 A の 2 次属性と呼ぶ. 概念ベースにおいて，「概念」は n 次までの属性の連鎖集合により定義されている．概念ベースの構造について図 1 に示す.

概念	属性／重み
雪	（雪／0.61），（白い／0.30），（下る／0.27），（結晶／0.25），…
白い	（雪／0.16），（白地／0.14），（色／0.14），（白髪／0.12），…
下る	（低い／0.23），（雪／0.21），（雨／0.20），（下がる／0.18），…

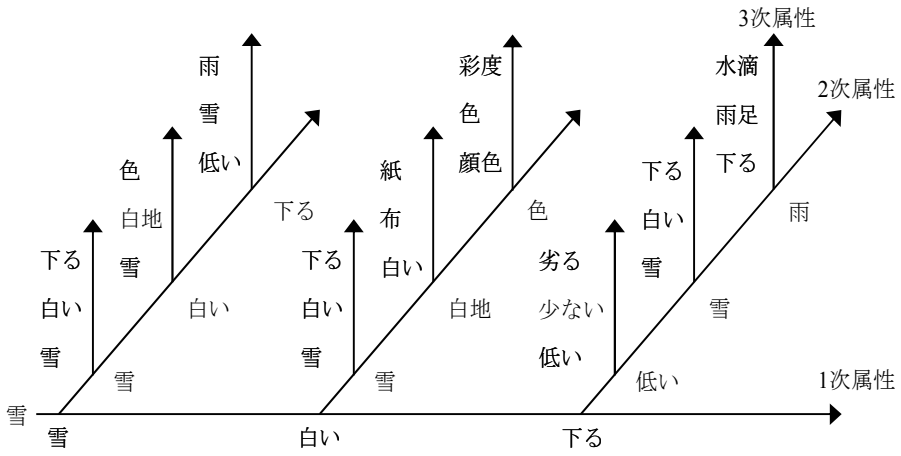

図 1：概念ベース(一部)

3.2 関連度計算

関連度計算方式[1]は，概念ベースに定義された語と語の関連の強さを，同義性，類似性のみに関わらず定量化する手法である．以下のような概念 A， B があるとする．

$$A = \{(a_1, w_1), (a_2, w_2), \cdots, (a_M, w_M)\} \quad (2)$$

$$B = \{(b_1, v_1), (b_2, v_2), \cdots, (b_N, v_N)\} \quad (3)$$

M , N は，それぞれ概念 A， B の属性数である．また，$a_i, w_i (1 \le i \le M)$ は，概念 A の属性とその重みである．同様に概念 B も $b_j, v_j (1 \le j \le N)$ で表される．2 次属性についても，以下のように定義される．

$$a_i = \{(a_{i1}, w_{i1}), (a_{i2}, w_{i2}), \cdots, (a_{im_k}, w_{im_k})\} \quad (4)$$

$$b_j = \{(b_{j1}, v_{j1}), (b_{j2}, v_{j2}), \cdots, (b_{jn_l}, v_{jn_l})\} \quad (5)$$

このとき，1 次属性 a_i と b_j の重み比率付き一致度 $DoM(a_i, b_j)$ は以下のように定義される．

$$DoM(a_i, b_j) = \sum_{a_{is} = b_{jt}} \min(w_{is}, v_{jt}) \quad (6)$$

重み比率付き一致度を 1 次属性全ての組合せに対して行い，一致度が大きいものから順に対応を決めていく．式(2)の概念 A に対して，一致度が最大となる組合せになるように，概念 B の属性を並べ替えたものを以下に示す.

$$B = \{(b_{x1}, v_{x1}), (b_{x2}, v_{x2}), \cdots (b_{xN}, v_{xN})\} \quad (7)$$

よって，これらの概念 A， B の関連度は次のようになる.

$$DoA(A, B) = \sum_i DoM(a_i, b_{xi}) \times (w_i + v_{xi}) / 2 \times (\min(w_i, v_{xi}) / \max(w_i, v_{xi})) \quad (8)$$

関連度の値は概念間の関連の強さを 0.0～1.0 の間の連続値で表す. 概念 A と B に対

して関連度を算出した例を表 1 に示す.

表 1：関連度計算の例

概念 *A*	概念 *B*	関連度の値
自動車	車	0.912
	飛行機	0.130
	学校	0.012

3.3 質問文意味理解システム

質問文意味理解システムは, 構文解析ツールを利用して, 質問文から質問対象語（質問文が求めている対象）とその条件（質問対象語にかかっている条件）を取得するシステムである．質問文意味理解システムでは，質問文中に疑問詞「誰」や「場所」の表現があった場合，質問対象語として「人物」,「場所」を獲得できる．また，疑問詞が「何」の場合や疑問詞がない場合でも質問対象語を獲得できる．質問文意味理解システムを用いて獲得した質問対象語の例を表 2 に示す.

表 2：質問対象語の獲得の例

質問文	質問対象語
同志社大学の前身である同志社英学校を創立したのは誰か？	人物
ドイツの首都はどこか？	場所
新島襄が創立に関わった大学は？	大学

質問文意味理解システムでは，約 90%の成功率で質問対象語を正しく獲得できると報告されており，優れたシステムといえる.

本研究では質問文意味理解システムを，質問対象語を獲得する手段として用いる.

4. Web を利用した技術

4.1 未定義語の属性獲得手法

未定義語（概念ベース未登録の概念）の属性獲得手法とは，未定義語の意味的特徴を表す属性（単語）とその重要性を表す重みの組を Web を用いて自動的に構成する手法である. この手法によりある単語に対して属性と重みを与えることを概念化と呼ぶ.

4.2 キーワードの意味分類体系ノードへの割付手法

キーワードの意味分類体系ノードへの割付手法とは，語の意味分類体系として定義されたノード（所属候補ノードとする）の中で，キーワードの所属するべきノードを提示する手法である.

4.3 TF・IDF

TF・IDF 法[8]とは，語の頻度と網羅性に基づいた重み付け手法である．TF はある文書中 *d* に出現する索引語 *t*（文書の内容を構成する要素）の頻度を表す尺度である．IDF はある索引語が全文書中のどれくらいの文書に出現するか（特定性）を表す尺度であり，式(9)で定義される．なお，*N* が検索対象となる文書集合中の全文書数，*df*(*t*) が索引語 *t* が出現する文書数である.

$$idf(t) = \log \frac{N}{df(t)} + 1 \tag{9}$$

4.4 Web・IDF

4.3 節で説明した IDF は一般的な文書（新聞や書籍など）を用いて索引語の特定性を考慮する手法である．一方，Web-IDF[6]は Web にある文書のみを用いて索引語の出現頻度を考慮する手法である．Web-IDF では式(9)の *N* を Google が保有している日本語のページ数，*df*(*t*)を索引語 *t* を Google で検索を行ったときのヒット件数とする．なお，Google は全言語において保有しているページ数は公開されているが，日本語のページとして保有している数は公開されていないため，日本語の文書として最も使われている「は」で検索を行ったヒット件数（1,980,000,000，2010 年 12 月 23 日現在）を Google が保有している日本語の全ページ数としている.

5. 質問対象語

質問対象語とは，質問文の求める回答の種類を表したものである．例えば，質問文「同志社の創立者は誰か？」の質問対象語は「人物」である．この例は「誰」という疑問詞から質問対象語を導いているが，文中の語を質問対象語として抜き出す場合もある．例えば,「新島襄が創立に関わった大学は？」という質問文の場合，質問対象語は「大学」となる．このように，質問対象語には質問文中から抜き出したものと，質問文の疑問詞から導かれたものがあり，回答スコア付けにおいてこれら 2 つの場合で処理を変更している.

6. 単語回答システム

本研究が提案する手法の流れは以下の通りである（図 2）．まず，質問文を入力した

後に，質問文解析を行い，質問応答のための検索条件（質問対象語および検索語）を決定する．次に，検索語を入力として Web から回答候補を獲得する．ここで，次処理の回答スコア付けを行うために検索語の概念化と回答候補の概念化を行う．最後に，回答候補に対して回答スコアを与える．スコアの上位順に順位付き回答候補リストとして出力する．

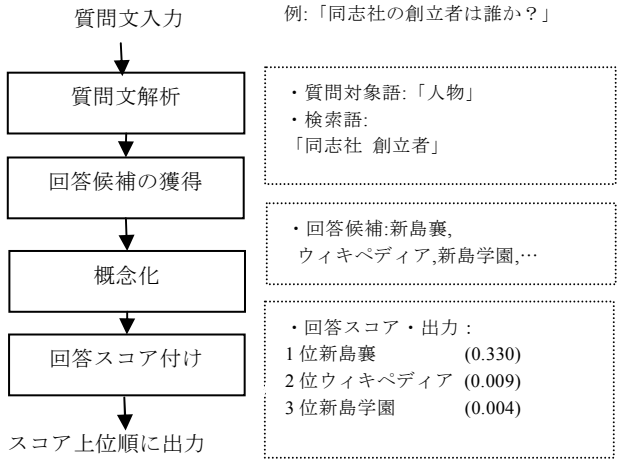

図 2：単語回答システムの流れ

6.1 質問文解析

・ 質問の質問対象語の決定

質問文意味理解システムにより暫定的な質問対象語を獲得し，質問対象語が「場所」の場合，検索語と「組織」および「場所」の関連度計算を調べ，関連度が大きいほうを質問対象語とする．質問対象語が「名前」および「名称」の場合は，文中から「名前」および「名称」の直前にある名詞を抽出し，質問対象語とする．それ以外の場合は暫定的な質問対象語をそのまま質問対象語とする．

・ 検索語の獲得

質問文に対して形態素解析ソフト「茶筌」[9]を用いて形態素解析を行い，自立語や複合語（数字やアルファベットの連続）を，質問文のキーワード（動詞・形容詞を除く）として抽出し，スペース区切りで繋げた語を検索語とする．以上の手順では，質問の内容をキーワードとして利用しているため，質問文が単文，複文に関係なく，あらゆる質問に対応することが可能であると考えられる．

6.2 回答候補の獲得

質問の回答候補を Web から獲得する．まず，検索語を入力として検索エンジンを用いて検索を行い，検索上位 100 件の検索結果ページの内容を取得する．次に，取得した文書群に対して，形態素解析を行い，自立語を抽出する．このとき，形態素に対して「名詞の連続は複合する」などの条件を用いて複合語を獲得する．最後に，抽出した自立語（複合語も含む）に対して，重み（主に頻度情報）を与え，重み上位 50 件を回答候補として獲得する．

6.3 概念化

未定義語の属性獲得手法を利用して，検索語および回答候補の概念化を行う．なお，検索語の概念化に利用する Web 文書は，質問文のキーワードから構成される検索語を入力として Web 検索を行い得られたものである．これにより，質問の回答のために必要な文書を網羅的に扱い，その文書の意味特徴を属性として獲得できる．概念化した検索語とは，質問の適合性を判断するための対象と捉えることができる．

6.4 回答スコア付け

回答候補が質問の答えとしてふさわしいかを判断するために，質問対象語に対する適合性および質問文に対する適合性の 2 つの観点からスコア付けを行う．本研究では，表 3 に示すスコア計算の積により回答スコアを与えている．

表 3：観点ごとのスコア計算方法

適合性の観点	スコア計算方法
質問文	意味得点
質問対象語	ノード得点

なお，ノード得点とは，質問対象語の種類に応じてノード動詞と Web の共起頻度を使い分け，より詳しく関係性を定量化する得点である．また，意味得点とは，検索語と回答候補の関連度と Web の共起頻度から算出する．

ノード得点と意味得点を用いることで，あらゆる固有名詞に対して，柔軟に質問対象語への適合性を評価（0～1 の値を取る）できる．

7. 評価

テストセットとしてアンケートにより収集した質問文 120 問を用いて，本研究で提案している手法の評価を行った．なお，テストセットの質問文には，正解となる単語（正答）を与えている．例を表 4 に示す．

表 4：テストセットの質問文の例

質問文	本州最北端の町はどこか？
質問対象語	場所
正答	青森県大間町

質問文の質問対象語の決定に関して，質問 106 文(88.3%)に対して質問対象語を正しく決定できた．決定方法は，三人の被験者に対し質問対象語が正しいかどうかアンケートを取り，全員が正しいと判断した場合○に，二人が正しいと判断した場合は△に，それ以外の場合は×とした．評価結果を図 3 に示す

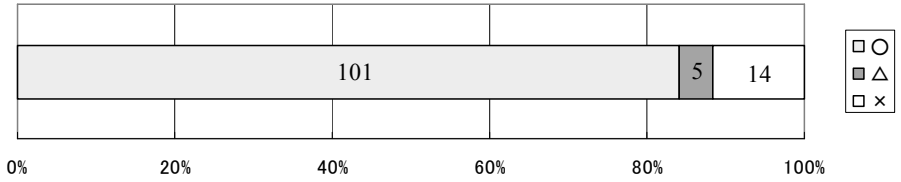

図 3：質問対象語の評価結果

また，回答候補の獲得において，質問文 120 問中 116 文(96.7%)に対して正答が獲得できた．

提案手法全体の評価結果として，図 4 に提案手法が出力したスコア付き回答候補リストに対して，正答が出現した順位の分布を示す．

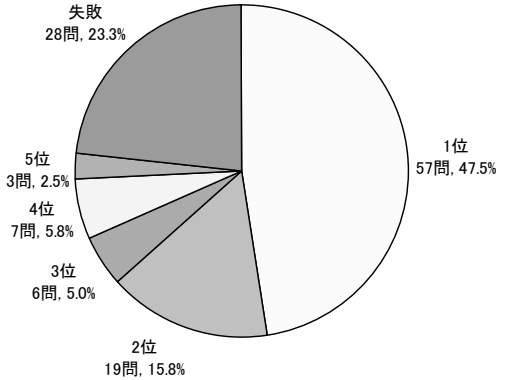

図 4：提案手法の出力における正答出現順位の分布

なお，図 4 における失敗とは，正答が 5 位以内に現れなかった場合である．出力 1 位に正答が出現した質問が 57 問となり，上位 5 位にかけて 92 問（76.7%）に対して正答が獲得できることがわかった．

8. 考察

まず，質問対象語の獲得について考察する．△や×の具体的な例を表 5 に示す．

表 5：△および×の例

質問文	質問対象語（出力）	判定
聖火リレーがはじめて行われたのはどこのオリンピックか？	場所	△
北アメリカ大陸に生息していた肉食恐竜は何か？	物	×
欧州連合の本部はどこにあるか？	組織	×

△と判定された質問対象語は，完璧ではないが間違ってもいないものが多かった．また，×と判定された質問文は大きく 2 つの場合に分けられる．1 つ目は，疑問詞「何」から質問対象語「物」を導いてしまった場合であり，2 つ目は疑問詞「どこ」から得た質問対象語「場所」を，「場所」か「組織」へ変換する際にうまく変換できなかった場合である．×と判定された 14 文の内，12 文がこの 2 つのどちらかであった．
そのため，今後質問対象語の精度を向上させるには，疑問詞「何」と「どこ」の扱いを変更する必要があると考えられる．

次に回答スコア付けに検索語と回答候補の共起ヒットを使用した場合の効果を確認するため提案手法と，検索語と回答候補の共起ヒットを使用しない手法を比較する．その結果，提案手法の MRR は比較手法より約 10%高くなった．回答スコア付けに検索語と回答候補の共起ヒットを利用した事による成功例として，質問文「イギリスの公用語は何か？」(正答：英語)が挙げられる．質問文「イギリスの公用語は何か？」に対する提案手法と比較手法の出力を表 6 に示す．Web 上に存在する「英語」が出現する文章のうち，「英語」が「イギリスの公用語」であると書かれた文章はごく一部である．そのため，通常の関連度計算のみでは「英語」は高い回答スコアを得ることができない．しかし，提案手法では検索語と回答候補の関連度について調べている．検索語「イギリス 公用語」が出現する文章には，高い確率で「英語」が出現しており，提案手法では正しい回答を出力することができた．

表 6：提案手法と比較手法の出力の例

	提案手法	比較手法
1 位	英語	アイルランド語
2 位	インド	ニュージーランド
3 位	英国	アメリカ英語
4 位	ニュージーランド	アイルランド
5 位	ロシア	オーストラリア

また，提案手法が正答を出力できなかった例として，質問文「聖火リレーが初めて行われたのはどこのオリンピックか？」(正答：ベルリンオリンピック)がある．質問文「聖火リレーが初めて行われたのはどこのオリンピックか？」に対する提案手法の出力を表 7 に示す．

表 7：提案手法の出力の例

	提案手法
1 位	五輪
2 位	北京
3 位	チベット問題
4 位	平和の祭典
5 位	アテネ
…	…
15 位	ベルリン

質問文「聖火リレーが初めて行われたのはどこのオリンピックか？」の検索語は，「聖火リレー 初めて オリンピック」となる．「初めて」は多くの文章に出現する単語であり，「聖火リレー」，「オリンピック」はオリンピックについて書かれた文章に多数出現する．そのため，過去のオリンピックである「ベルリンオリンピック」について書かれた文書よりも，最近のオリンピックである「北京オリンピック」について書かれた文章が多く取得されてしまった．このような失敗の原因は，質問文から検索後を取得する際に情報が欠落しているためと考えられる．例えば，質問文「聖火リレーが初めて行われたのはどこのオリンピックか？」を検索語「聖火リレー 初めて オリンピック」に変換した結果，質問文には存在した「聖火リレー」と「初めて」の意味のつながりが失われている．

この問題を解決するためには，質問文から検索語を取得する際の意味の欠落をなくす必要がある．そのため，今後は質問文からの検索語の取得方法を改良する必要がある．

9. おわりに

本研究では，質問文に対する単語回答を Web から獲得する，質問文の種類に応じた web 検索による単語回答システムを提案した．

結果として，質問文 120 問に対して，出力 5 位以内に正答が存在する割合（76.7%）で正答を選択することに成功し，多くの質問文に対して正答を柔軟に選択できることを示した．

謝辞 本研究の一部は，科学研究費補助金（若手研究（B）21700241）の補助を受けて行った．

参考文献

1) 源明和也，渡部広一，河岡司，“単語解答を求める複雑な質問文を対象とした知的 Web 検索方式”，情報処理学会研究報告. ICS，117-122，2009.
2) 奥村紀之，土屋誠司，渡部広一，河岡司，“概念間の関連度計算のための大規模概念ベースの構築”，自然言語処理，Vol.14，No.5，pp.41-64，2007.
3) 渡部広一，奥村紀之，河岡司，“概念の意味属性と共起情報を用いた関連度計算方式”，自然言語処理，Vol.13，No.1，pp.53-74，2006.
4) 古川成道，渡部広一，河岡司，“概念ベースを用いた知的検索における曖昧な質問文の意味理解”，第 18 回人工知能学会全国大会論文集，2D1-10，2004.
5) Google， http://www.google.co.jp/
6) 辻泰希，渡部広一，河岡司，“www を用いた概念ベースにない新概念およびその属性獲得手法”，第 18 回人工知能学会全国大会論文集，2D1-01，2004.
7) 後藤和人，土屋誠司，渡部広一，河岡司，“Web を用いた未知語検索キーワードのシソーラスノードへの割付け手法”，自然言語処理，Vol.15，No.3，pp.91-113，2008.
8) 徳永健伸，“言語処理と計算 5 情報検索と言語処理”，東京大学出版会，1999.
9) 奈良先端科学技術大学院大学，http://chasen-legacy.sourceforge.jp/