社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2009-ICS-154（8）
2009/3/2

語の意味処理に基づく知識文からの回答抽出方式

東 宏樹† 吉村 枝里子‡ 渡部 広一‡ 河岡 司‡

†‡同志社大学 〒610-0394 京都府京田辺市多々羅都谷 1-3

E-mail: †dth0704@mail4.doshisha.ac.jp, ‡{eyoshimura, hwatabe, tkawaoka}@indy.doshisha.ac.jp

あらまし 人間とのコミュニケーションを可能とするコンピュータの実現のためには，コンピュータにも人間と同じような判断能力が望まれる．人間同士の日常生活の中には，ある人間が質問し，相手の人間がその質問に対して正解を判断し，回答することができる．そこで，本稿で提案する手法は，単語の意味に着目して，語の意味から質問の正解の可能性があるかを判断し，回答語の絞込みを行う．このように，質問文の意味を理解し，知識文から質問の正解となる語を決定する手法を提案する。

キーワード 意味処理，質問回答，教養知識

A Method of Getting Answer by Processing Words Meaning from Knowledge Sentences

Hiroki AZUMA† Eriko YOSHIMURA‡ and Hirokazu WATABE‡ and Tsukasa KAWAOKA‡

†‡Graduate School of Engineering, Doshisha University

1-3 Miyakodani Tatara Kyotanabe-shi, Kyoto, 610-0394 Japan

E-mail: †dth0704@mail4.doshisha.ac.jp, ‡{eyoshimura, hwatabe, tkawaoka}@indy.doshisha.ac.jp

Abstract A similar judgment ability to man is hoped for to the computer for the achievement of the computer that man and communications are possible. Man associates the answer candidate with the question thrown by the interlocutor, and can find the correct answer. It equips the computer with this judgment mechanism. The technique for proposing it by this text pays attention to the meaning of the word, judges whether there is a possibility of the correct answer of the question from the word meaning, and narrows the answer candidate. The meaning of the question sentence is understood from this technique, and it proposes the technique for deciding the word that becomes a correct answer of the question from the knowledge sentence.

Keyword Meaning Processing , Question-Answering, Educational Knowledge

1. はじめに

近年，コンピュータが急速に発展し，人間が生活する中で必要不可欠なものとなってきている．そこで，人間からの要求や意図を理解し，人間とコミュニケーションが取れるコンピュータが望まれる．

人間とのコミュニケーションを可能とするコンピュータの実現のためには，コンピュータにも人間と同じような判断能力が望まれる．日常生活には，「江戸幕府を開いた人物は誰ですか？」という質問に対して，答えを知らなくても「徳川家康は江戸に幕府を開いた将軍です.」という知識文を与えられれば，質問の意味を理解し，この知識文の情報にその正解が存在するかどうかを判断して，「徳川家康です.」と回答することができる．しかし，コンピュータには知識文が与えられていてもどの語が答えなのかと判断することが難しい．そこで，本稿で提案する手法は，単語の意味に着目して，語の意味から質問の正解の可能性があるかを判断し，回答語の絞込みを行う．このように単語の意味に着目して回答語の絞込みを行うことを「語の意味処理」と定義し，それによって，質問に対する回答語を決定する．

2. 研究概要

まず，本研究で対象とする文について説明する．質問文については，「名詞」を問う質問が質問の中で最も多いと考えられるので，名詞を問う質問文とする．そのため，疑問詞が「Who」，「When」，「Where」，「What」または「疑問詞なし」の質問文を対象とする．なお，「疑問詞なし」の質問文とは，「世界最長の河川は？」といったような「何ですか？」を省略した形のことであり，種類としては「What」と同じ種類とする．次に，知識文は，単文，重文，複文を問わず，句点が一つしか存在しない知識文を対象とする．また，様々な質問応答の内容の中で，後述する教養知識[1]を対象とし，既存の教養知識ベース[1]を利用する．その中でも，歴史，地理に関する教養知識ベースを利用し，質問文と知識文を歴史，地理に関する文とする．

次に，1 章で述べたような質問応答を可能とするために，様々な手法を用いて回答語を抽出する方式を考

案した．質問文の入力から回答語の決定までの流れを図1に示し，説明する．

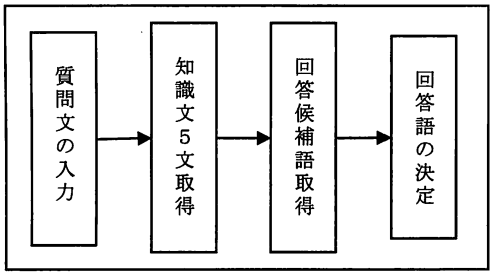

図1：システム全体の流れ

まず，質問文を入力し，後述する知識文抽出システム[1]を利用して質問文と関連の高い知識文を5文取得する．そして，取得した5文から形態素解析，シソーラス[2]，教養シソーラス[1]を用いて構築した「単一文フレーム」と既存システムである「質問文意味理解システム[3]」を利用して質問文に対する「回答候補語」を取得する．さらに，この「回答候補語」から後述する「概念ベース[4]」を利用することで，最終的な「回答語」を決定する．

以上のような流れで処理を行う．その際に利用する連想メカニズム，シソーラスについて，以下に説明する．また本研究で扱う教養知識についても説明する．

2.1. 連想メカニズム

人間は知識にない表現についても，知識にある語から連想することで対応することができる．例えば，『列車』という言葉を知らなくても『車』を知っていれば，「車の一種」だと予想することができる．このようなメカニズムを，連想メカニズムと呼ぶ．

コンピュータは知識ベースにない語については柔軟な対応を行うことができない．そこで，概念ベースや関連度計算といった関連技術を用いることで連想メカニズムを組み込み，様々な語に対応する．

2.1.1. 概念ベース

概念ベースとは，複数の国語辞書などから機械的に構築された知識ベースである．概念ベースには約12万語の概念が格納されており，総属性数は約254万語である．概念Aをその概念の意味を表す属性a_iと属性の重要性を表す重みw_iの対の集合として定義すると，以下のように表される．

$A = \{(a_1, w_1), (a_2, w_2), \ldots, (a_m, w_m)\}$

Ex）雪 = {(氷,0.5), (白い,0.3), (降る,0.17), ... }

2.1.2. 関連度計算

関連度[5]とは，任意の概念と概念の関連の強さを概念ベースの属性と重みを用いて定量化した0以上1以下の数値のことであり，関連度が1になるのは二つの概念が同一のときだけである．従って，関連度は概念間の関連が強いほど1に近い値を取り，弱いほど0に近い値を取る．このように関連の強さを定量化することで，概念間の関連の強弱を判断することができる．表1に例を示す．

表1：関連度計算の例

基準概念	対象概念	関連度
飛行機	航空機	0.66
	自動車	0.12
	電車	0.06

2.2. シソーラス

シソーラスとは，一般名詞を広く同義，類義的に整理したもので，約2700個の意味属性（ノード）の上位・下位関係，全体・部分関係が木構造で示され，ノードに属する名詞として約13万語（リーフ）が登録されている辞書である．

2.3. 教養知識

教養知識とは，小・中学校で習う程度の知識の中で比較的頻繁に出現する知識とする．例えば「徳川家康は1603年に江戸幕府を開いた」のような知識を指す．教養知識を確立するために，入出力として必要な文章形式の教養知識ベース，そして，教養知識に関する語の連想を可能とする教養概念ベース[2]，また，語と語の同義類義関係，上位下位をツリー構造で表した教養シソーラスの三つの知識を構築している．

2.3.1. 教養知識ベース

教養知識ベースは教養4科目である，「地理」，「歴史」，「音楽」，「美術」に関する質問文に対して，回答となる文章を持たせ返答することを目的として構築された．そこで，教養知識ベースの知識文を回答として用いる．

教養知識ベースに格納する教養知識文の定義を行う．まず，「地理」については，地球上の固定された三次元空間にある事物及びそれに関する事象と定義する．次に，「歴史」の定義を行う．「歴史」とは，広く一般に知られている知識・事柄のうち，過去に発生した出来事や人物，物の名前などであり，事実であると考えられているものとする．次に，「音楽」の定義を行う．「音楽」とは，時間的に規則性があるなど，ランダムではない特性を持ち，かつ人間が楽しむことのできる雑音以外の音とする．また，このような特性を持つ音を様々な方法で発したり，聴いたり，想像したり，それに合わせて体を動かしたりしてたのしむ行為のことも音楽とする．最後に「美術」の定義を行う．「美術」とは，絵画・彫刻・建築・工芸などの，視覚的，空間的な美を表現する造形芸術の総称と定義する．これら教養4科目は全て小学・中学校で習う範囲とほぼ同じ範疇にある．

また，教養知識ベースの知識文は，「地理」，「歴史」，「音楽」，「美術」に関する教科書などに比較的頻繁に出現する文章と定義する．それぞれ手作業で入力し，地理 561 文，歴史 916 文，音楽 725 文，美術 398 文である．

本研究では，「地理」と「歴史」の質問文を対象としているため，「地理」と「歴史」の教養知識ベースの知識文を用いる．

2.3.2. 教養概念ベース

教養知識に関して関連度計算を行う際，教養 4 科目に関する固有名詞（徳川家康，ゴッホ）等は概念ベースに登録されていない為，関連度が算出されない．そこで概念ベースに登録されていない語（未定義語）についての関連度計算をするためにはこの未定義語が概念ベースに登録されている必要がある．そこで，概念ベースに教養 4 科目に関する未定義語を追加し，教養 4 科目に関する固有名詞に関しても関連度を算出できるようにする為，未定義語の概念ベースへの追加を行う．以後，概念ベースを既存の概念ベース，概念ベースに教養 4 科目に関する未定義語を追加した概念ベースを教養概念ベースと呼ぶ．

2.3.3. 教養シソーラス

教養シソーラスは 2.2 節で説明したシソーラスを参考に，教養 4 科目に関する包含関係を人手で知識として構築したものである．ノードは 162 語，リーフは約 4000 語登録されている．

3. 知識文の取得

本章から，提案手法の流れに沿って説明する．

まず，質問文を入力し，知識文抽出システムを用いて教養知識ベースから知識文を 5 文取得する，教養知識ベースの知識文は，「地理」，「歴史」，「音楽」，「美術」に関する教科書などに比較的頻繁に出現する文章と定義し，それぞれ手作業で入力した．そして，その知識文と質問文の関連性を関連度計算方式[2]によって取得し，関連の高い上位 5 文の知識文を取得する．知識文取得の例を図 2 に示す．

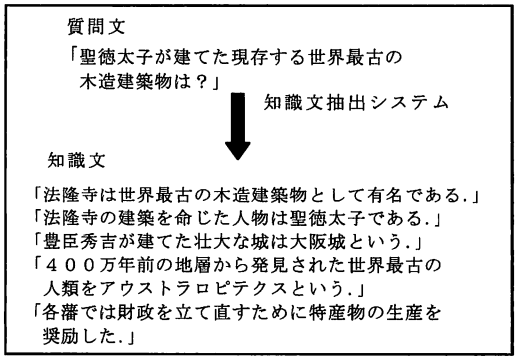

図 2：知識文の取得の例

ここで，知識文抽出システムを説明する．

3.1. 知識文抽出システム

知識文抽出システムは，教養に関する問題かどうかを判定する判定部と，教養に関する問題に対して知識文を出力する回答部に分けられる．判定部は，自然言語の質問文に対して，それが教養に関する問題かどうかを判断する処理部分であり，回答部は教養に関すると判断された問題に対して正解が存在すると考えられる知識文 5 文を取得する部分である．教養に関する知識文抽出システムについての概要は図 3 の通りである．

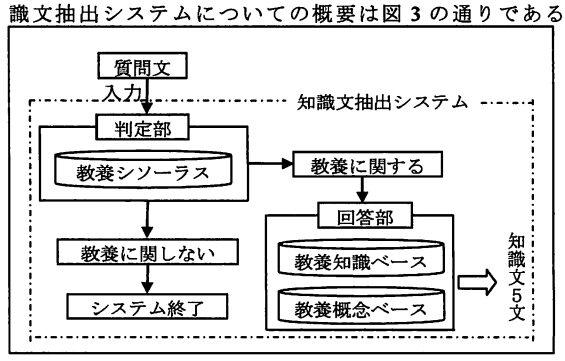

図 3 知識文抽出システム

処理の流れとしては，まず，図 3 上部の知識文抽出システムの判定部で自然言語の質問文が教養に関する問題なのかどうかの判定を行う．教養シソーラスのリーフが質問文に存在すれば教養に関すると判定する．そして，教養に関しないと判定されたときは回答部の処理にかけずにシステム終了とし，教養に関するときは，回答部にて教養概念ベースと教養知識ベースを利用することにより質問文と教養知識ベースの質問文に関連があると判断できる知識文すべてに対して，関連度計算を行う．そして，関連度が高い上位 5 文を質問文の正解が含まれる知識文として抽出する．

4. 回答候補語の取得

本章では質問文から取得した知識文から回答候補語を取得する部分を説明する．その際，知識文中の名詞ごとに意味を判別させるために「単一文フレーム」を構築した．また，質問文から疑問詞などを取得するために，後述する「質問文意味理解システム」を利用した．以下に既存の「質問文意味理解システム」と構築した「単一文フレーム」の説明をし，回答候補語の絞込み方法と取得方法を説明する．

4.1. 質問文意味理解システム

質問文意味理解システムとは，質問文から質問対象語とその条件，疑問詞を取得するシステムである．質問対象語とは，質問文が答えを求めている語のことである．例えば，「明日電車で行く場所はどこですか？」という質問文の場合，質問対象語は「場所」となり，その条件は「明日電車で行く」となる．そして疑問詞は

「Where」を取得する．

4.2. 単一文フレーム

単一文フレームとは，一つの文を対象として，文中に存在する名詞を意味別のフレームに格納するものである．例えば「徳川家康は1603年に江戸幕府を開いた．」といった知識文を，表2のように名詞の意味を判別して整理された状態にフレーム分けする．

表2：単一文フレームへの格納例

人物	時間	場所	その他	教養
徳川家康	1603年		江戸幕府(組織)	徳川家康

フレームの項目は「人物」，「時間」，「場所」，「その他」，「教養」とする．その他フレームに関しては，「キーワード」を作成し，意味を詳細化している．教養フレームに関しては，今回，教養に関する知識文を対象としているために，形態素解析やシソーラスでは対応できない用語を格納するために「教養シソーラス」を用いて格納する．

キーワードとはある用語を端的に説明する，用語と関連のある適切なシソーラスノードである．質問文の疑問詞が「What」もしくは「疑問詞なし」で与えられた場合，質問文が何を質問しているのかの判断が難しいが，このキーワードを付与することで，回答語を絞り込みやすくすることが可能となる．表1の格納例では「江戸幕府(組織)」の「(組織)」がそれに当たる．このようにキーワードを設定し，知識文中の名詞から，シソーラス内で検索を行い，その名詞のキーワードを決める．設定したキーワードは「出来事」，「制度」，「法律」，「組織」，「言語」，「宗教」，「未知語」7つとし，すべてその他フレーム内の語とする．

格納方法を説明する．図4に流れを示し，流れに沿って以下に説明する．

まず知識文を形態素解析し，「名詞」もしくは「連続した名詞」を取得する．そこで，「品詞の詳細」に「人名」や「地域名」，と表示されていれば，それぞれ「人物」フレーム，「場所」フレームに格納する．

「品詞の詳細」にそのような表示が無い場合は「シソーラス」を利用してキーワードを検索する．これは「取得した名詞の語尾」に着目して，例えば「～事件」，「～憲法」といった表記であれば，その語の親ノードをシソーラスで検索する．この例の場合，それぞれ「出来事」と「法律」といった，キーワードを検索できるため，そのキーワードを付随させて，「その他」フレームへ格納する．

最後に，再び形態素解析の結果の「品詞の詳細」に「固有名詞」が存在すれば，キーワードを付随させずに「その他」フレームへ格納する．また，「品詞の詳細」が「一般名詞」であっても5語以上の名詞の連続であった場合は，固有名詞とみなし，「その他」フレームへ格納する．つまり，名詞を取得した際に「一般名詞」の「単語の名詞」場合は，回答語になる可能性はきわめて低いと判断し，フレームへは格納しないことにする．

「時間」フレームに関しては，「～年」，「～時」，「～月」，「～日」，「～週」，「～世紀」，「～ごろ」，「～初頭」，「～中期」，「～後期」，「～時代」の全11語の名詞を時間独特な言い回しと設定し，語尾が表記一致すれば「時間」フレームへ格納する．

また，「教養」フレームに関しては，知識文中に教養シソーラス内のリーフに存在する語があれば，その語を「教養」フレームへ格納する．

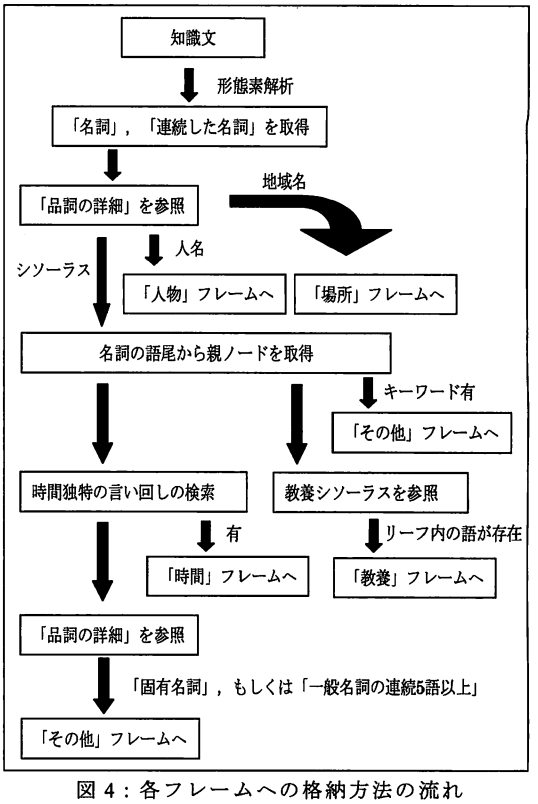

図4：各フレームへの格納方法の流れ

このようにして，知識文中に存在する名詞を単一文フレームへ格納することにする．これにより，知識文の名詞を意味別に区別することができる．

4.3. 回答候補語の絞込み

取得した知識文5文のうち，1文ずつから回答候補語を取得する．その際，単一文フレームを用いて回答候補語の絞込みを行う．

まず，取得した知識文を単一文フレームへ格納する．そして，質問文の疑問詞によって，回答候補語を絞り込む．疑問詞が「Who」の場合は，単一文フレームの

「人物」フレームに存在する語に絞る．同様に，「When」の場合は「時間」フレーム，「Where」の場合は「場所」フレームに絞る．「What」もしくは「疑問詞なし」の場合は，疑問詞では絞込みが難しいため，質問文から取得した，「質問対象語」を利用して絞込みを行う．

「質問対象語」をシソーラス内で検索，もしくは，「質問対象語」の親ノードをシソーラスから取得し，「質問対象語」もしくは「取得した親ノード」に，単一文フレームで設定した「キーワード」が存在するかを検索する．キーワードが発見できれば，単一文フレーム内のそのキーワードが付随している語に絞る．キーワードが発見できなければ，「教養」フレーム内の語も含み，フレーム内のすべての語を回答候補語とする．

このような処理を，知識文 5 文に対して行い，5 文の中から回答の可能性がある語，つまり回答候補語を取得する．

この流れを図 5 に示す．

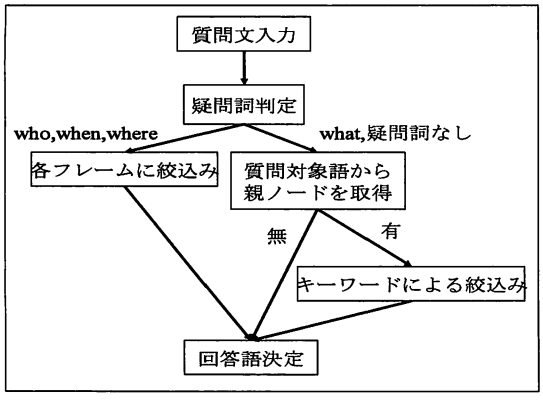

図 5：回答語候補取得の流れ

質問文を 1 文，知識文を 1 文として，例を図 6 に示し，以下に説明する．

知識文「第一次世界大戦の後，革命のおこったドイツで制定された国民主権の憲法をワイマール憲法という」
質問文「第一次世界大戦後にドイツで制定した憲法は何ですか？」

単一文フレーム

人物	時間	場所	その他	教養
		ドイツ	第一次世界大戦(出来事) ワイマール憲法(法律)	ドイツ ワイマール憲法

疑問詞「what」　質問対象語「憲法」

回答候補語　「ワイマール憲法」

図 6：回答候補語取得の例

この場合，質問文の疑問詞が「What」のためフレームでの絞り込みができない．よって，シソーラスより質問文の質問対象語「憲法」の親ノードを取得する．「憲法」の親ノードとして「法律,掟,制度,抽象物...」と取得でき,キーワードの「法律」が存在するため，「ワイマール憲法（法律）」が取得でき，回答候補語に「ワイマール憲法」が取得できる．このような処理を，取得した知識文 5 文に対して行い，複数の回答候補語を取得する．

5. 回答語の決定

本章では，回答候補語から，最終的に回答する回答語の決定を行う．そこで，教養概念ベースを利用する．

まず，質問文を形態素解析し，名詞と形容詞を取得する．この際，「誰」や「何」のような疑問詞に関する名詞を含む語は除外する．例を示す．

質問文:「第一次世界大戦の後にドイツで制定された憲法を何という？」

取得語:「第一次大戦」，「後」，「ドイツ」，「制定」，「憲法」

このように質問文から語を取得し，その語が教養概念ベース内で回答候補語の属性に何語存在するかを検索する．質問文に対して正しい回答であれば，質問文中の語が属性として多く存在していると考えられるからである．この例の場合，回答候補語には「ワイマール憲法」と「日本国憲法」が取得できる．図 7 の様に，「ワイマール憲法」には「憲法」と「ドイツ」が属性に存在し，「日本国憲法」には「憲法」のみ属性に存在した．このことから，この質問文に対しての回答としては「ワイマール憲法」の方が正解であると判断できる．

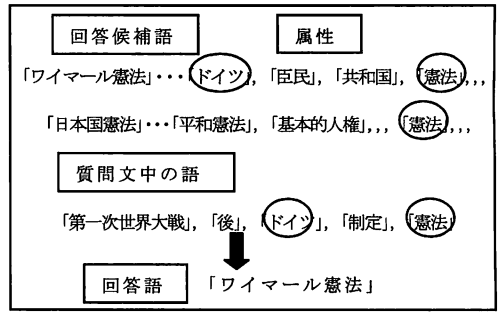

図 7：回答語の決定の例

次に，属性の数が同じ回答候補語が存在した場合について説明する．知識文を取得する際に知識文抽出システムを用いている．このとき，質問文と知識文の関連度を取得している．属性の数が同じだった場合はこの関連度が高い文から取得した回答候補語を回答語として採用する．高い関連度で取得したということは，質問文と知識文の内容が近いと判断できるからである．

また，回答候補語の中には，未定義語，つまり教養概念ベースに存在しない語が少し存在する．そのような語に関しては未定義語の属性獲得手法[6]を用いるこ

とで属性を取得し，この手法を可能にする．

6. 評価と考察

本稿では，語の意味に基づいた「単一文フレーム」を用いた処理を提案し，質問文と関連の高い知識文 5 文から回答を導いた．本手法を用いて導いた回答が正解であるかを評価し，考察を加える．

6.1. 評価

評価データとしては歴史に関する質問文 50 文と地理に関する質問文 50 文の計 100 文の質問文を「小中学校にて習う地理，歴史の一般教養について」というアンケートにより作成した．知識文はそれらの質問文を入力して知識文抽出システムで得られた各々5 文を利用した．

評価方法としては出力された回答語が質問文の正解であれば○，不正解であれば×とし，すべて目視で行った．なお，取得した知識文 5 文の中に質問文の正解である回答語が含まれているとは限らない．

評価結果を図 8 に示し，成功例を図 9 に示す．

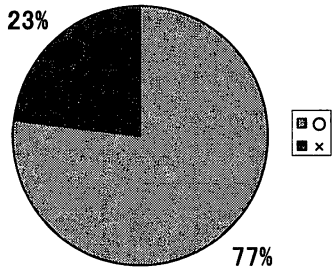

図 8：評価結果

(表記例)「質問文」・・・「回答語」
(成功例)
「徒然草を書いた人は誰ですか?」・・・「吉田兼好」
「東大寺南大門にある彫刻は何という像ですか?」・・・「金剛力士像」
「中国の首都はどこの都市か?」・・・「北京」
「ドイツを流れている国際河川は何という川ですか?」・・・「ライン川」

図 9：成功例

評価結果は精度 77%となり，提案手法に有効性があることがわかる．また，取得した知識文 5 文の中に質問文の正解が含まれていない例は，作成した質問文 100 文中 3 文存在した．

6.2. 考察

失敗例を挙げて考察する．「室町幕府を滅ぼした武将は誰ですか？」という質問文が入力されたときの回答候補語は「足利尊氏」，「織田信長」，「新田義貞」，「後醍醐天皇」，「足利義満」を取得する．質問文の正解としては「織田信長」となるが，このとき，5 章で説明した手法で回答語を決定すると，回答語は「後醍醐天皇」となる．これは，質問文中の語である「武将」が「後醍醐天皇」の属性にのみ存在し，他の回答候補語の属性の中には質問文中の語が存在しなかったからである．この問題の改善には，「足利尊氏」や「織田信長」といった語の属性に「武将」を入れる必要があり，教養概念ベースの精錬が望まれる．また，この失敗例の場合，「足利尊氏」が質問文と最も関連の高い知識文から取得した回答候補語であるため，教養概念ベースを精錬しても「足利尊氏」が回答語となる．つまり，この失敗を改善するには質問文中に存在する「滅ぼした」という語に着目する必要があると考えられる．このように，質問文中の語に重要度を考慮に入れ，属性を検索する必要がある．この重要度を考慮に入れれば，回答候補語と質問文との関連がより定量的に表現でき，回答語を決定する判断が可能になると考えられる．

7. 終わりに

本稿では，入力した質問文から，その質問文と関連の高い知識文を取得し，単一文フレームを構築し，それを利用することで，質問文の正解となる語を決定する回答抽出方式を提案した．コンピュータに知識を持たせれば，単一文フレームを利用して名詞の意味を判別し，教養概念ベースを利用して連想することから的確な回答語を抽出し回答することができた．しかし，まだこの方式では様々な失敗例が存在し，6.2 節の考察を基に，質問文の正解を回答する確率を上げるように改善する必要がある．

今回は教養知識を対象として実験を行ったが，時事情報や会話文にも対応させていくことで，より人間と柔軟なコミュニケーションが可能になると考えられる．

文　献

[1] 中本一志，渡部広一，河岡司，“web 情報文からの教養知識の自動学習方式”，信学技報，NLC2007-93，pp.33-37，Feb.2007.

[2] NTT コミュニケーション科学研究所監修，“日本語語彙体系”，岩波書店，東京，1997.

[3] 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”，人工知能学会全国大会，2D1-10，June.2004.

[4] 小島一秀，渡部広一，河岡司，“連想システムのための概念ベース構成法－属性信頼度の考え方に基づく属性重みの決定”，自然言語処理，Vol.9，No.5，pp.93-110，Oct.2002.

[5] 渡部広一，河岡司，“常識的判断のための概念間の関連度評価モデル”，自然言語処理，Vol.8，No.2，pp.39-54，April.2001.

[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”，人工知能学会全国大会，2D1-01，June.2004.